

Please type a plus sign (+) inside this box ☐

+

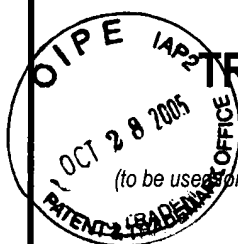
AF/1647 IFW

PTO/SB/21 (6-99)

Approved for use through 09/30/2000. OMB 0651-0031  
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

10-31-05

 <h1>TRANSMITTAL FORM</h1>		Application Number	09/997,614
		Filing Date	NOVEMBER 15, 2001
		First Named Inventor	DAVID BOTSTEIN
		Group/Art Unit	1647
		Examiner Name	WEGERT, SANDRA L.
Total Number of Pages in This Submission	220	Attorney Docket Number	39780-2730 P1C29

## ENCLOSURES (check all that apply)

<input type="checkbox"/> Fee Transmittal Form <input type="checkbox"/> Fee Attached <input type="checkbox"/> Amendment / Response <input type="checkbox"/> After Final <input type="checkbox"/> Version With Markings Showing Changes <input type="checkbox"/> Affidavits/declaration(s) <input type="checkbox"/> Extension of Time Request <input type="checkbox"/> Information Disclosure Statement <input type="checkbox"/> Certified Copy of Priority Document(s) <input type="checkbox"/> Response to Missing Parts/ Incomplete Application <input type="checkbox"/> Response to Missing Parts under 37 CFR 1.52 or 1.53 <input type="checkbox"/> Copy of Notice	<input type="checkbox"/> Copy of an Assignment <input type="checkbox"/> Drawing(s) <input type="checkbox"/> Licensing-related Papers <input type="checkbox"/> Petition Routing Slip (PTO/SB/69) and Accompanying Petition <input type="checkbox"/> Petition to Convert to a Provisional Application <input type="checkbox"/> Power of Attorney, by Assignee to Exclusion of Inventor Under 37 C.F.R. §3.71 With Revocation of Prior Powers <input type="checkbox"/> Terminal Disclaimer <input type="checkbox"/> Small Entity Statement <input type="checkbox"/> Request for Refund	<input type="checkbox"/> After Allowance Communication to Group <input type="checkbox"/> Appeal Communication to Board of Appeals and Interferences <input checked="" type="checkbox"/> <b>APPEAL COMMUNICATION TO GROUP (APPEAL NOTICE, BRIEF, REPLY BRIEF)</b> <input type="checkbox"/> Proprietary Information <input type="checkbox"/> Status Letter <input checked="" type="checkbox"/> <b>ADDITIONAL ENCLOSURE(S) (PLEASE IDENTIFY BELOW):</b> <input checked="" type="checkbox"/> <b>EVIDENCE APPENDIX ITEMS 1-19; and POSTCARD</b>
Remarks <b>AUTHORIZATION TO CHARGE DEPOSIT ACCOUNT 08-1641 FOR ANY FEES DUE IN CONNECTION WITH THIS PAPER, REFERENCING ATTORNEY'S DOCKET NO. 39780-2730 P1C29.</b>		

## SIGNATURE OF APPLICANT, ATTORNEY OR AGENT

Firm or Individual name	HELLER EHRMAN LLP		DAPHNE REDDY (Reg. No. 53,507)	
	275 Middlefield Road, Menlo Park, California 94025	Telephone: (650) 324-7000	Facsimile: (650) 324-0638	
Signature	<i>Daphne Reddy</i>			
Date	OCTOBER 28, 2005	Customer Number:	35489	

## CERTIFICATE OF EXPRESS MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 C.F.R. §1.10 on the date indicated below and addressed to: **MAIL STOP APPEAL BRIEF - PATENTS**, Commissioner for Patents, PO Box 1450, Alexandria, Virginia 22313-1450, on this date: **OCTOBER 28, 2005**

Express Mail Label **EV 582 623 533 US**

Typed or printed name	C. FONG		
Signature	<i>C. Fong</i>	Date	OCTOBER 28, 2005

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Mail Stop \_\_\_\_, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

BEST AVAILABLE COPY

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

David BOTSTEIN, et al.

Application Serial No. 09/997,614

Filed: November 15, 2001

For: **SECRETED AND TRANSMEMBRANE  
POLYPEPTIDES AND NUCLEIC  
ACIDS ENCODING THE SAME**

) Examiner: Wegert, Sandra

) Art Unit: 1647

) Confirmation No: 7398

) Attorney's Docket No. 39780-2730 P1C29

) Customer No. 35489

**EXPRESS MAIL LABEL NO. : EV 582 623 533 US**

**DATE MAILED: OCTOBER 28, 2005**

**ON APPEAL TO THE BOARD OF PATENT APPEALS AND INTERFERENCES**

**APPELLANTS' BRIEF**

**MAIL STOP APPEAL BRIEF - PATENTS**

Commissioner for Patents

P.O. Box 1450

Alexandria, Virginia 22313-1450

Dear Sir:

This Appeal Brief, filed in connection with the above captioned patent application, is responsive to the Final Office Action mailed on October 20, 2004. A Notice of Appeal was filed herein on January 20, 2005. This brief is timely filed since August 20, 2005 is a Saturday. A request for a five month extension of time is filed concurrently herewith. Appellants hereby appeal to the Board of Patent Appeals and Interferences from the final rejection in this case.

The Commissioner is authorized to charge any fees which may be required, including extension fees, or credit any overpayment to Deposit Account No. **08-1641** (referencing Attorney's Docket No. **39780-2730 P1C29**).

The following constitutes the Appellants' Brief on Appeal.

**I. REAL PARTY IN INTEREST**

The real party in interest is Genentech, Inc., South San Francisco, California, by an assignment of the parent application, U.S. Serial No. 09/941,992 recorded November 16, 2001, at Reel 012176 and Frame 0450.

**II. RELATED APPEALS AND INTERFERENCES**

The claims pending in the current application are directed to a polypeptide referred to herein as "PRO1097". There exist two related patent applications, (1) U.S. Serial No. 09/997,628, filed November 15, 2001 (containing claims directed to antibodies to the PRO1097 polypeptide), and (2) U.S. Serial No. 09/989,723, filed November 19, 2001 (containing claims directed to nucleic acids encoding PRO1097 polypeptides). These two related applications are also under final rejection from the same Examiner and based upon the same outstanding rejection, therefore appeal of these final rejections are being pursued independently and concurrently herewith.

**III. STATUS OF CLAIMS**

Claims 119-126 and 129-131 are in this application.

Claims 1-118 and 127-128 have been canceled.

Claims 119-126 and 129-131 stand rejected and Appellants appeal the rejection of these claims.

A copy of the rejected claims in the present Appeal is provided as Appendix A.

**IV. STATUS OF AMENDMENTS**

In an Amendment filed on March 10, 2005 after the mailing of the Final Office of October 18, 2004, a request under Rule C.F.R. §1.48 for correction of inventorship was filed, and this amendment was entered for purposes of this appeal.

## **V. SUMMARY OF CLAIMED SUBJECT MATTER**

The invention claimed in the present application is related to an isolated polypeptide comprising the amino acid sequence of the polypeptide of SEQ ID NO: 349, referred to in the present application as "PRO1097." The PRO1097 gene was shown for the first time in the present application to be significantly amplified in human lung or colon cancers as compared to normal, non-cancerous human tissue controls (Example 170). This feature is specifically recited in claim 124, and carried by all claims dependent from claim 124. In addition, the invention also claims the amino acid sequence of the polypeptide of SEQ ID NO: 349, lacking its associated signal-peptide; or the amino acid sequence of the polypeptide encoded by the full-length coding sequence of the cDNA deposited under ATCC accession number 203044 (Claim 124-126 and 129). The invention is further directed to polypeptides having at least 80%, 85%, 90%, 95% or 99% amino acid sequence identity to the amino acid sequence of the polypeptide of SEQ ID NO: 349; the amino acid sequence of the polypeptide of SEQ ID NO: 349, lacking its associated signal peptide; or the amino acid sequence of the polypeptide encoded by the full-length coding sequence of the cDNA deposited under ATCC accession number 203044, wherein the nucleic acid encoding said polypeptide is amplified in lung or colon tumor (Claims 119-123). The invention is further directed to a chimeric polypeptide comprising one of the above polypeptides fused to a heterologous polypeptide (Claim 130), and to a chimeric polypeptide wherein the heterologous polypeptide is an epitope tag or an Fc region of an immunoglobulin (Claim 131).

The amino acid sequence of the native "PRO1097" polypeptide and the nucleic acid sequence encoding this polypeptide (referred to in the present application as "DNA59841-1460") are shown in the present specification as SEQ ID NOs: 349 and 348, respectively, and in Figures 244 and 243, respectively found on pages 299, lines 30-34. A full-length PRO1097 polypeptide having the amino acid sequence of SEQ ID NO:349 is described in the specification at, for example, on pages 218-220, line 30 onwards and the isolation of cDNA clones encoding PRO1097 of SEQ ID NO:349 is described in Example 107, page 489 of the specification. The specification discloses that various portions of the PRO1097 polypeptide possess significant sequence similarity to the glycoprotease family of proteins and the acyltransferase ChoActase/COT/CPT family (see, for example, page 218, lines 31-34).

PRO polypeptide variants having at least about 80-99% amino acid sequence identity with a full length PRO polypeptide sequence, or a PRO polypeptide sequence lacking the signal



peptide are described in the specification at, for example, page 305, line 23 onwards, and percent amino acid sequence identity determination is described at, for example, pages 306-308, line 14 onwards. The preparation of chimeric PRO polypeptides (claims 130 and 131), including those wherein the heterologous polypeptide is an epitope tag or an Fc region of an immunoglobulin, is set forth in the specification at page 374, lines 24 to page 375, line 9. Examples 140-143 and page 376, line 12 onwards describe the expression of PRO polypeptides in various host cells, including *E. coli*, mammalian cells, yeast and Baculovirus-infected insect cells.

Finally, Example 170, in the specification at page 539, line 19, to page 555, line 5, sets forth a 'Gene Amplification assay' which shows that the PRO1097 gene is amplified in the genome of certain human lung or colon cancers (see Table 9, page 550). The profiles of various primary lung and colon tumors used for screening the PRO polypeptide compounds of the invention in the gene amplification assay are summarized on Table 8, page 546 of the specification.

#### **VI. GROUND S OF REJECTION TO BE REVIEWED ON APPEAL**

1. Whether Claims 119-126 and 129-131 should be accorded priority of provisional Application 60/141,037, filed 19 November, 2001.
2. Whether Claims 119-126 and 129-131 satisfy the utility/ enablement requirement under 35 U.S.C. §101/112, first paragraph.
3. Whether Claims 119-126 and 129-131 satisfy the written description requirement under 35 U.S.C. §112, first paragraph.

#### **VII. GROUPING OF CLAIMS**

With respect to Issue 1, all claims (Claims 119-126 and 129-131) stand and fall together.  
With respect to Issue 2, all claims (Claims 119-126 and 129-131) stand and fall together.  
Issue 3 concerns only Claims 119-126, which claims stand and fall together.

## VIII. ARGUMENTS

### Summary of the Arguments

#### Issue 1: Priority

The instant application has not been granted the earlier priority date on the grounds that “although disclosing the same experimental assays as the instant specification, do not enable the instant invention and therefore do not impart utility...”

Appellants submit that data derived from the Gene Amplification assay was first disclosed in U. S. Application Serial No. 60/141,037, filed 19 November, 2001 for the gene encoding the claimed PRO1097 polypeptide. Appellants further submit that, the same detailed reasons discussed below under the section on Issue II: Utility/ Enablement, are sufficient to also establish patentable utility for U. S. Application Serial No. 60/141,037. Hence, Appellants should be able to rely upon this provisional application to provide an effective filing date of 19 November, 2001 for the instant application.

#### Issue 2: Utility/ Enablement

Claims 119-126 and 129-131 stand rejected under 35 U.S.C. §101/ 112, first paragraph as allegedly lacking either a specific and substantial asserted utility or a well established utility. Appellants have previously submitted that patentable utility for the PRO1097 polypeptides is based upon the gene amplification data for the gene encoding the PRO1097 polypeptide. The specification discloses that the gene encoding PRO1097 showed significant amplification, ranging from 2.313 to 2.346 fold in two different lung primary tumors and 2.114 to 2.532 fold in three different colon primary tumors. Therefore, such a gene is useful as a marker for the diagnosis of cancer, and for monitoring cancer development and/or for measuring the efficacy of cancer therapy.

In the first Office action mailed May 3, 2004, the Examiner cited references Hittelman *et al.* and Crowell *et al.*, to show that "a slight increase in clone numbers in a cancerous tissue is no doubt due to an increased number of chromosomes, a very common characteristic of cancerous and non-cancerous epithelial cells." Appellants submit that, in fact, the Hittelman reference supports the Appellants position that there is utility in identifying genetic biomarkers in epithelial tissues at cancer risk (see Hittelman, abstract, line 4-7).

The Examiner further cited references Skolnick *et al.*, Bork *et al.*, Doerks *et al.*, Hesselgesser *et al.* and Blease *et al.* to show that "function cannot be predicted based solely on structural similarity to a protein found in sequence databases." Appellants had argued in their response of August 3, 2004 that Appellants assertion for utility of PRO1097 was not based on structural similarity.

The Examiner further asserted on page 3 of the Final Office Action mailed October 18, 2004 that amplification of the PRO1097 polynucleotide does not impart a specific, substantial, and credible utility to the PRO1097 polypeptide since, "there is no evidence regarding whether or not PRO1097 mRNA or polypeptide levels are also increased in (these) cancer." In support of this assertion, the Examiner cited references by Pennica *et al.*, Haynes *et al.* and Hu *et al.*

Appellants submit that, the teachings of Pennica *et al.* are not directed towards genes in general but to a single gene or genes within a single family and thus, their teachings cannot support a general conclusion regarding correlation between gene amplification and mRNA or protein levels. Further, Appellants submit that the teachings of Haynes *et al.* in fact, meets the "more likely than not standard" and shows that a positive correlation exists between mRNA and protein. And based on the nature of the statistical analysis performed in one class of genes in Hu *et al.*, the Examiner's conclusions are not reliably supported. Thus, Appellants submit that these references do not conclusively establish a *prima facie* case for lack of utility.

In contrast, Appellants have submitted ample evidence to show that, in general, if a gene is amplified in cancer, it is more likely than not that the encoded protein will be expressed at an elevated level. First, the articles by Orntoft *et al.*, Hyman *et al.*, and Pollack *et al.* (made of record in Appellants' Response filed July 7, 2004) collectively teach that in general, gene amplification increases mRNA expression. Second, the Declaration of Dr. Paul Polakis (made of record in Appellants' Response filed August 3, 2004), principal investigator of the Tumor Antigen Project of Genentech, Inc., the assignee of the present application, shows that, in general, there is a correlation between mRNA levels and polypeptide levels. Appellants further note that the sale of gene expression chips to measure mRNA levels is a highly successful business, with a company such as Affymetrix recording 168.3 million dollars in sales of their GeneChip arrays in 2004. Clearly, the research community believes that the information obtained from these chips is useful (i.e., that it is more likely than not informative of the protein level).

Taken together, although there are some examples in the scientific art that do not fit within the central dogma of molecular biology that there is a correlation between DNA, mRNA, and polypeptide levels, these instances are exceptions rather than the rule. In the majority of amplified genes, as exemplified by Orntoft *et al.*, Hyman *et al.*, Pollack *et al.*, the Polakis Declaration and the widespread use of array chips, the teachings in the art overwhelmingly show that gene amplification influences gene expression at the mRNA and protein levels. Therefore, one of skill in the art would reasonably expect in this instance, based on the amplification data for the PRO1097 gene, that the PRO1097 polypeptide is concomitantly overexpressed. Thus, the claimed PRO1097 polypeptides also have utility in the diagnosis of cancer.

Appellants further submit that even if there is no correlation between gene amplification and increased mRNA/protein expression, (which Appellants expressly do not concede), a polypeptide encoded by a gene that is amplified in cancer would still have a specific, substantial, and credible utility. Appellants submit that, as evidenced by the Ashkenazi Declaration and the teachings of Hanna and Mornin (both made of record in Appellants' Response filed August 3, 2004), simultaneous testing of gene amplification and gene product over-expression enables more accurate tumor classification, even if the gene-product, the protein, is not over-expressed. This leads to better determination of a suitable therapy for the tumor, as demonstrated by a real-world example of the breast cancer marker HER-2/neu. Accordingly, Appellants submit that when the proper legal standard is applied, one should reach the conclusion that the present application discloses at least one patentable utility for the claimed PRO1097 polypeptides.

The Examiner also cited references Smith *et al.*, and Brenner *et al.*, to support an enablement rejection that "it is not predictable which amino acids are necessary to maintain the functional characteristics of protein".

Appellants submit that, besides the detailed protocol for the gene amplification assay, the specification further provides ample guidance to allow the skilled artisan to identify those polypeptides which meet the limitations of the claims, including, how to identify polypeptides based on % identity to (SEQ ID NO: 349), how to make PRO1097 polypeptides, etc. Prediction of the amino acid(s) necessary for functionality is not necessary to practice the invention. Appellants further submit that, based on the gene amplification data and the substantial, credible, asserted utility of PRO1097 polypeptides in the diagnosis of lung or colon cancer, one of

ordinary skill would know exactly how to make and use these claimed polypeptides for the diagnosis of cancers, without any undue experimentation.

### Issue 3: Written Description

Claims 119-126 and 129-131 stand rejected under 35 U.S.C. §112, first paragraph, allegedly "as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.." (Page 9 of the Final Office Action mailed October 18, 2004).

The factors to be considered in evidencing possession of a claimed genus include "disclosure of complete or partial structure, physical and/or chemical properties, functional characteristics, structure/function correlation, methods of making the claimed product, or any combination thereof." Appellants note that the claims recite structural features, namely, 80% sequence identity to SEQ ID NO:349, which are common to the genus. The genus of claimed polypeptides is further defined by having a specific functional activity for the encoding nucleic acids, namely, that the encoding nucleic acid is amplified in lung or colon tumors. The specification provides detailed guidance as to how to identify polypeptides having at least 80% amino acid sequence identity to SEQ ID NO:349 (PRO1097), as well as detailed protocols for determining whether a gene encoding a variant PRO1097 protein is amplified in lung or colon or tumor. Thus one of skill in the art could easily identify whether a variant PRO1097 sequence falls within the parameters of the claimed invention.

Accordingly, a description of the claimed genus has been achieved by the recitation of both structural and functional characteristics.

### Response to Rejections

#### **ISSUE 1. U.S. Provisional Application No. 60/141,037 Satisfies the Utility Requirement of 35 U.S.C. § 101/ § 112, First Paragraph based on the results of the Gene Amplification assay**

Appellants have asserted that U.S. Provisional Application No. 60/141,037, filed November 19, 2001, discloses the gene amplification assay (shown in Example 170 of the instant specification) and establishes patentable utility for the claimed PRO1097 polypeptides.

Appellants submit, for the reasons set forth below under Issue 2 for Utility/ Enablement, that the results of the gene amplification assay disclosed in the specification of U.S. Application No. 60/141,037, provides at least one credible, substantial and specific asserted utility for the claimed PRO1097 polypeptides under 35 U.S.C. §101/§112, first paragraph. Accordingly, Appellants respectfully request that the subject matter of the instant claims be granted the November 19, 2001, priority date of U.S. Provisional Application No. 60/141,037.

**ISSUE 2. Claims 119-126 and 129-131 are supported by a credible, specific and substantial asserted utility, and thus meet the utility requirement of 35 U.S.C. § 101/ 112, first paragraph**

The sole basis for the Examiner's rejection of Claims 119-126 and 129-131 under this section is that the data presented in Example 170 of the present specification is allegedly insufficient under the present legal standards to establish a patentable utility under 35 U.S.C. § 101 for the presently claimed subject matter.

Claims 119-126 and 129-131 stand further rejected under 35 U.S.C. §112, first paragraph, allegedly "since the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility for the reasons set forth above, one skilled in the art clearly would not know how to use the claimed invention."

Appellants strongly disagree and, therefore, respectfully traverse the rejection.

**A. The Legal Standard For Utility Under 35 U.S.C. § 101**

According to 35 U.S.C. § 101:

Whoever invents or discovers any new and *useful* process, machine, manufacture, or composition of matter, or any new and *useful* improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title. (Emphasis added.)

In interpreting the utility requirement, in *Brenner v. Manson*<sup>1</sup> the Supreme Court held that the *quid pro quo* contemplated by the U.S. Constitution between the public interest and the interest of the inventors required that a patent applicant disclose a "substantial utility" for his or

---

<sup>1</sup> *Brenner v. Manson*, 383 U.S. 519, 148 U.S.P.Q. (BNA) 689 (1966).

her invention, i.e. a utility "where specific benefit exists in currently available form."<sup>2</sup> The Court concluded that "a patent is not a hunting license. It is not a reward for the search, but compensation for its successful conclusion. A patent system must be related to the world of commerce rather than the realm of philosophy."<sup>3</sup>

Later, in *Nelson v. Bowler*<sup>4</sup> the C.C.P.A. acknowledged that tests evidencing pharmacological activity of a compound may establish practical utility, even though they may not establish a specific therapeutic use. The court held that "since it is crucial to provide researchers with an incentive to disclose pharmaceutical activities in as many compounds as possible, we conclude adequate proof of any such activity constitutes a showing of practical utility."<sup>5</sup>

In *Cross v. Iizuka*<sup>6</sup> the C.A.F.C. reaffirmed *Nelson*, and added that *in vitro* results might be sufficient to support practical utility, explaining that "*in vitro* testing, in general, is relatively less complex, less time consuming, and less expensive than *in vivo* testing. Moreover, *in vitro* results with the particular pharmacological activity are generally predictive of *in vivo* test results, i.e. there is a reasonable correlation there between."<sup>7</sup> The court perceived "No insurmountable difficulty" in finding that, under appropriate circumstances, "in vitro testing, may establish a practical utility."<sup>8</sup>

---

<sup>2</sup> *Id.* at 534, 148 U.S.P.Q. (BNA) at 695.

<sup>3</sup> *Id.* at 536, 148 U.S.P.Q. (BNA) at 696.

<sup>4</sup> *Nelson v. Bowler*, 626 F.2d 853, 206 U.S.P.Q. (BNA) 881 (C.C.P.A. 1980).

<sup>5</sup> *Id.* at 856, 206 U.S.P.Q. (BNA) at 883.

<sup>6</sup> *Cross v. Iizuka*, 753 F.2d 1047, 224 U.S.P.Q. (BNA) 739 (Fed. Cir. 1985).

<sup>7</sup> *Id.* at 1050, 224 U.S.P.Q. (BNA) at 747.

<sup>8</sup> *Id.*

The case law has also clearly established that Appellants' statements of utility are usually sufficient, unless such statement of utility is unbelievable on its face.<sup>9</sup> The PTO has the initial burden to prove that Appellants' claims of usefulness are not believable on their face.<sup>10</sup> In general, an Applicant's assertion of utility creates a presumption of utility that will be sufficient to satisfy the utility requirement of 35 U.S.C. §101, "unless there is a reason for one skilled in the art to question the objective truth of the statement of utility or its scope."<sup>11, 12</sup>

Compliance with 35 U.S.C. §101 is a question of fact.<sup>13</sup> The evidentiary standard to be used throughout *ex parte* examination in setting forth a rejection is a preponderance of the totality of the evidence under consideration.<sup>14</sup> Thus, to overcome the presumption of truth that an assertion of utility by the applicant enjoys, the Examiner must establish that it is more likely than not that one of ordinary skill in the art would doubt the truth of the statement of utility. Only after the Examiner made a proper *prima facie* showing of lack of utility, does the burden of rebuttal shift to the applicant. The issue will then be decided on the totality of evidence.

The well established case law is clearly reflected in the Utility Examination Guidelines ("Utility Guidelines")<sup>15</sup>, which acknowledge that an invention complies with the utility requirement of 35 U.S.C. §101, if it has at least one asserted "specific, substantial, and credible utility" or a "well-established utility." Under the Utility Guidelines, a utility is "specific" when

---

<sup>9</sup> *In re Gazave*, 379 F.2d 973, 154 U.S.P.Q. (BNA) 92 (C.C.P.A. 1967).

<sup>10</sup> *Ibid.*

<sup>11</sup> *In re Langer*, 503 F.2d 1380, 1391, 183 U.S.P.Q. (BNA) 288, 297 (C.C.P.A. 1974).

<sup>12</sup> *See also In re Jolles*, 628 F.2d 1322, 206 USPQ 885 (C.C.P.A. 1980); *In re Irons*, 340 F.2d 974, 144 USPQ 351 (1965); *In re Sichert*, 566 F.2d 1154, 1159, 196 USPQ 209, 212-13 (C.C.P.A. 1977).

<sup>13</sup> *Raytheon v. Roper*, 724 F.2d 951, 956, 220 U.S.P.Q. (BNA) 592, 596 (Fed. Cir. 1983) cert. denied, 469 US 835 (1984).

<sup>14</sup> *In re Oetiker*, 977 F.2d 1443, 1445, 24 U.S.P.Q.2d (BNA) 1443, 1444 (Fed. Cir. 1992).

<sup>15</sup> 66 Fed. Reg. 1092 (2001).



it is particular to the subject matter claimed. For example, it is generally not enough to state that a nucleic acid is useful as a diagnostic without also identifying the conditions that are to be diagnosed.

In explaining the “substantial utility” standard, M.P.E.P. §2107.01 cautions, however, that Office personnel must be careful not to interpret the phrase “immediate benefit to the public” or similar formulations used in certain court decisions to mean that products or services based on the claimed invention must be “currently available” to the public in order to satisfy the utility requirement. “Rather, any reasonable use that an applicant has identified for the invention that can be viewed as providing a public benefit should be accepted as sufficient, at least with regard to defining a ‘substantial’ utility.”<sup>16</sup> Indeed, the ‘Guidelines for Examination of Applications for Compliance With the Utility Requirement,’<sup>17</sup> gives the following instruction to patent examiners: “If the applicant has asserted that the claimed invention is useful for any particular practical purpose . . . and the assertion would be considered credible by a person of ordinary skill in the art, do not impose a rejection based on lack of utility.”

#### **B. Proper Application of the Legal Standard**

Appellants respectfully submit that the data presented in Example 170 starting on page 539 of the specification of the specification and the cumulative evidence of record, which underlies the current dispute, indeed support a “specific, substantial and credible” asserted utility for the presently claimed invention.

Example 170 describes the results obtained using a very well-known and routinely employed polymerase chain reaction (PCR)-based assay, the TaqMan<sup>TM</sup> PCR assay, also referred to herein as the gene amplification assay. This assay allows one to quantitatively measure the level of gene amplification in a given sample, say, a tumor extract, or a cell line. It was well known in the art at the time the invention was made that gene amplification is an essential mechanism for oncogene activation. Appellants isolated genomic DNA from a variety of primary cancers and cancer cell lines that are listed in Table 9 (pages 539 onwards of the specification),

---

<sup>16</sup> M.P.E.P. §2107.01.

<sup>17</sup> M.P.E.P. §2107 II (B)(1).

including primary lung and colon cancers of the type and stage indicated in Table 8 (page 546). The tumor samples were tested in triplicates with Taqman<sup>TM</sup> primers and with internal controls, beta-actin and GADPH in order to quantitatively compare DNA levels between samples (page 548, lines 33-34). As a negative control, DNA was isolated from the cells of ten normal healthy individuals, which was pooled and used as a control (page 539, lines 27-29) and also, no-template controls (page 548, lines 33-34). The results of TaqMan<sup>TM</sup> PCR are reported in  $\Delta C_t$  units, as explained in the passage on page 539, lines 37-39. One unit corresponds to one PCR cycle or approximately a 2-fold amplification, relative to control, two units correspond to 4-fold, 3 units to 8-fold amplification and so on. Using this PCR-based assay, Appellants showed that the gene encoding for PRO1097 was amplified, that is, it showed approximately 1.21- 1.23  $\Delta C_t$  units for lung tumors and 1.08-1.34  $\Delta C_t$  units for colon tumors which corresponds to  $2^{1.21}$  -  $2^{1.23}$ - fold amplification in lung and  $2^{1.08}$  -  $2^{1.34}$ - fold amplification in colon tumors respectively, or **2.313 to 2.346 fold** in two different lung primary tumors and **2.114 to 2.532 fold** in three different colon primary tumors.

The Examiner acknowledged that there was an “increase” in DNA, but stated that the increase was “slight” or “small”. In fact, based on Hittelman *et al.*, the Examiner stated that such a “slight increase in clone numbers in cancerous tissue is no doubt due to an increased number of chromosomes, a very common characteristic of cancerous and non-cancerous epithelial cells.” Appellants disagree.

Hittelman studied premalignant lesions and suggests that epithelial tumors develop through a multistep process driven by genetic instability (see abstract). Hittelman showed that a subset of the same molecular changes found in associated tumor were also found in premalignant lesions, suggesting that these premalignant lesions might represent precursor lesions for associated tumors, i.e., a manifestation of a multistep tumorigenesis process. (See Hittelman, page 4, last three lines). Appellants therefore submit that, contrary to the Examiner’s rejection, the Hittelman reference strongly supports the Appellants position that there is utility in identifying genetic biomarkers in epithelial tissues at cancer risk (also see Hittelman, abstract, line 4-7). Hittelman adds on page 2, fourth paragraph, line 3 that “it is important to identify individuals at significantly increased cancer risk who might best benefit from different types of intervention”. Taken together, even if Appellants were to show that the observed PRO1097 gene amplification were due to chromosomal aneuploidy (which Appellants do not contend to),

identifying genetic biomarkers like the PRO1097 gene with this aneuploidy is a very important and useful step, according to Hittelman, in identifying individuals at significantly increased cancer risk. Therefore, Hittelman supports at least one utility for the PRO1097 gene, that is, as a genetic biomarker for cancer or precancerous cells. As one skilled in the art would clearly know, early detection of lung cancer provides information in advance about risk assessment, prognosis and therapy for lung cancer.

As evidence that the “increase in DNA” in the gene amplification assay is significant, Appellants submit a Declaration by Dr. Audrey Goddard (copy enclosed herewith). The Declaration by Dr. Audrey Goddard provides a statement by an expert in the relevant art that “fold amplification” values of at least 2-fold are considered significant in the TaqMan™ PCR gene amplification assay. This Declaration is necessary at this time to counter the assertion that the gene amplification data does not have utility. The issue whether the fold increase in the gene amplification assay for the PRO1097 gene was “significant” was not raised in the First Office action mailed May 3, 2004 nor in the Final Office action mailed October 18, 2004. Therefore, this declaration addressing “significance” was not presented earlier since the Appellants had no opportunity or reason to address this issue until now. Thus good and sufficient reasons exist why this Declaration is necessary and was not earlier presented. Appellants therefore submit that entry of the Goddard Declaration is appropriate at this time and respectfully request that it be considered..

Appellants particularly draw the Board's attention to page 3 of the Goddard Declaration which clearly states that:

It is further my considered scientific opinion that an at least **2-fold increase** in gene copy number in a tumor tissue sample relative to a normal (*i.e.*, non-tumor) sample is significant and useful in that the detected increase in gene copy number in the tumor sample relative to the normal sample serves as a basis for using relative gene copy number as quantitated by the TaqMan PCR technique as a diagnostic marker for the presence or absence of tumor in a tissue sample of unknown pathology. Accordingly, a gene identified as being amplified at least 2-fold by the quantitative TaqMan PCR assay in a tumor sample relative to a normal sample is **useful as a marker for the diagnosis of cancer**, for monitoring cancer development and/or for measuring the efficacy of cancer therapy. (Emphasis added).

Accordingly, the 2.313 to 2.346 fold in two different lung primary tumors and 2.114 to 2.532 fold in three different colon primary tumors would be considered significant and credible by one skilled in the art, based upon the facts disclosed in the Goddard Declaration.

Further Appellants submit that the fact that two lung tumor samples and three colon tumor samples tested positive in this study does not make the gene amplification data, by any means, less significant or spurious. As any skilled artisan in the field of oncology would easily appreciate, not all tumor markers are generally associated with every tumor, or even, with most tumors. In fact, some tumor markers are useful for identifying rare malignancies. That is, the association of the tumor marker with a particular type of tumor lesion may be rare, or, the occurrence of that particular kind of tumor lesion itself may be rare. In either event, even these rare tumor markers, which may not give a positive hit with most common tumors, have great value in tumor diagnosis, and consequently, in tumor prognosis. The skilled artisan would know that such tumor markers are very useful for better classification of tumors. Therefore, whether the PRO1097 gene is amplified in two lung/ three colon tumors or in most tumors is not relevant to its identification as a tumor marker, or its patentable utility. Rather, whether the amplification data for PRO1097 is significant is what lends support to its usefulness as a tumor marker. It was well known in the art at the time of filing of the application that gene amplification, which occurs in most solid tumors like lung and colon cancers, is generally associated with poor prognosis. Therefore, the PRO1097 gene becomes an important diagnostic marker to identify such malignant lung or colon cancers, even if the malignancy associated with PRO1097 molecule is a rare occurrence. Accordingly, the present specification clearly discloses enough evidence that the gene encoding the PRO1097 polypeptide is significantly amplified in certain types of lung or colon tumors and is therefore, a valuable diagnostic marker for identifying certain types of lung or colon cancers.

On page 4 of the final Office Action, the Examiner points out that “there is no evidence regarding whether or not PRO1097 mRNA or polypeptide levels are also increased in this cancer”. The Examiner points out, especially on page 4-5 of the Final Office Action mailed on October 18, 2004, that:

"what is often seen is a lack of correlation between DNA amplification and increased peptide levels (Pennica *et al.*) As discussed by Haynes *et al.*, polypeptide levels cannot be accurately predicted from mRNA levels...the literature cautions researches against

drawing conclusions based on small changes in transcript expression levels between normal and cancerous tissue."

Appellants strongly disagree. Appellants submit that the Examiner applied an improper legal standard when making this rejection. The evidentiary standard to be used throughout *ex parte* examination of a patent application is a preponderance of the totality of the evidence under consideration. Thus, to overcome the presumption of truth that an assertion of utility by the applicant enjoys, the Examiner must establish that it is more likely than not that one of ordinary skill in the art would doubt the truth of the statement of utility. Only after the Examiner has made a proper *prima facie* showing of lack of utility, does the burden of rebuttal shift to the applicant.

Accordingly, it is not a legal requirement to establish a necessary correlation between an increase in the copy number of the DNA and protein expression levels that would correlate to the disease state or that it is imperative to find evidence that DNA amplification is "necessarily" or "always" associated with overexpression of the gene product. Appellants respectfully submit that when the proper evidentiary standard is applied, a correlation must be acknowledged.

First of all, the teachings of Pennica *et al.* are specific to *WISP* genes, a specific class of closely related molecules. Pennica *et al.* showed that there was good correlation between DNA and mRNA expression levels for the *WISP-1* gene but not for *WISP-2* and *WISP-3* genes. But, the fact that in the case of closely related molecules, there seemed to be no correlation between gene amplification and the level of mRNA/protein expression does not establish that it is more likely than not, in general, that such correlation does not exist. As discussed above, the standard is not absolute certainty. Pennica *et al.* has no teaching whatsoever about the correlation of gene amplification and protein expression for genes in general. Indeed, the working hypothesis among those skilled in the art is that, if a gene is amplified in cancer, the encoded protein is likely to be expressed at an elevated level. In fact, as noted even in Pennica *et al.*, "[a]n analysis of *WISP-1* gene amplification and expression in human colon tumors *showed a correlation between DNA amplification and over-expression . . .*" (Pennica *et al.*, page 14722, left column, first full paragraph, emphasis added). Accordingly, Appellants respectfully submit that Pennica *et al.* teaches nothing conclusive regarding the absence of correlation between gene amplification and over-expression of mRNA or polypeptides in most genes, in general.

Further, contrary to the Examiner's reading, the Haynes *et al.* reference teaches that "there was a *general trend but no strong correlation* between protein [expression] and transcript levels" (Emphasis added). For example, in Figure 1, there is a positive correlation between mRNA and protein levels amongst most of the 80 yeast proteins studied. In fact, very few data points deviated or scattered away from the expected normal and no data points showed a negative correlation between mRNA and protein levels (i.e. an increase in mRNA resulted in a decrease in protein levels). The analysis by Haynes *et al.* is not relevant to the current application. Haynes was studying yeast cells and not human cells. Haynes *et al.* notes that their analysis focused on the 80 most abundant proteins in the yeast lysate (page 1867). Haynes *et al.* states "since many important regulatory protein are present only at low abundance, these would not be amenable to analysis" (page 1867). Further, Haynes *et al.* compared the protein expression levels of these naturally abundant proteins to mRNA expression levels from published SAGE frequency tables. (page 1863) Accordingly, Haynes *et al.* did not compare mRNA expression levels and protein levels in the same yeast cells. And thus, the analysis by Haynes *et al.* is not applicable to the present application. In fact, when the proper legal standard is used, Haynes' teachings clearly support the Appellants' position and is all that's needed to meet the "more likely than not" evidentiary standard. Again, accurate prediction is not the standard.

The Examiner further cited Hu *et al.*, to show that "by the current literature...one skilled in the art would not assume that a small increase in gene copy number would correlate with significantly increased mRNA or polypeptide levels" (Page 5 of the Final Office action mailed October 18, 2004).

First of all, as discussed above, the increase in DNA copy number for the PRO1097 gene is significant. Further, Appellants respectfully submit that, contrary to the Examiner's assertion, the cited Hu *et al.* reference does not conclusively establish a *prima facie* case for lack of utility for the PRO1097 molecule. The Hu *et al.* reference is entitled "Analysis of Genomic and Proteomic Data using Advanced Literature Mining" (emphasis added). Therefore, as the title itself suggests, the conclusions in this reference are based upon statistical analysis of information obtained from published literature, and not from experimental data. Hu *et al.* performed statistical analysis to provide evidence for a relationship between mRNA expression and biological function of a given molecule (as in disease). The conclusions of Hu *et al.* however,

only apply to a specific type of breast tumor (estrogen receptor (ER)-positive breast tumor) and cannot be generalized to breast cancer genes in general, let alone to cancer genes in general. Interestingly, the observed correlation was only found among ER-positive (breast) tumors not ER-negative tumors.” (See page 412, left column).

Moreover, the analytical methods utilized by Hu *et al.* have certain statistical drawbacks, as the authors themselves admit. For instance, according to Hu *et al.*, “*different statistical methods*” were applied to “*estimate* the strength of gene-disease relationships and evaluated the results.” (See page 406, left column, emphasis added). Using these different statistical methods, Hu *et al.* “[a]ssessed the relative strengths of gene-disease relationships based on the frequency of both co-citation and single citation.” (See page 411, left column). As is well known in the art, different statistical methods allow different variables to be manipulated to affect the resulting outcome. In this regard, the authors disclose that, “Initial attempts to search the literature ” using the list of genes, gene names, gene symbols, and frequently used synonyms generated by the authors “revealed several sources of false positives and false negatives.” (See page 406, right column). The authors add that the false positives caused by “duplicative and unrelated meanings for the term” were “difficult to manage.” Therefore, in order to minimize such false positives, Hu *et al.* disclose that these terms “had to be eliminated entirely, thereby reducing the false positive rate but unavoidably under-representing some genes.” *Id.* (emphasis added). Hence, Hu *et al.* had to manipulate certain aspects of the input data, in order to generate, in their opinion, meaningful results. Further, because the frequency of citation for a given molecule and its relationship to disease only reflects the current research interest of a molecule, and not the true biological function of the molecule, as the authors themselves acknowledge, the “[r]elationship established by frequency of co-citation do not necessarily represent a true biological link.” (See page 411, right column). Therefore, based on these findings, the authors add, “[t]his may reflect *a bias in the literature* to study the more prevalent type of tumor in the population. Furthermore, this emphasizes that caution must be taken when interpreting experiments that may contain subpopulations that behave very differently.” *Id.* (Emphasis added). In other words, some molecules may have been underrepresented merely because they were less frequently cited or studied in literature compared to other more well-cited or studied genes. Therefore, Hu *et al.*’s conclusions are not based on genes/mRNA *in general*.

Therefore, Appellants submit that, based on the nature of the statistical analysis performed herein, and in particular, based on Hu's analysis of *one* class of genes, namely, the estrogen receptor (ER)-positive breast tumor genes, the conclusions drawn by the Examiner, namely that, "genes displaying a 5-fold change or less (mRNA expression) in tumors compared to normal showed no evidence of a correlation between altered gene expression and a known role in the disease (in general)" is not reliably supported.

Therefore, when the proper legal standard is used, a *prima facie* case of lack of utility has not been met based on the cited references Pennica *et al.*, Haynes *et al.* or Hu *et al.* by the Examiner.

On the contrary, Appellants submit that Example 170 in the specification further discloses that, "(a)mplification is associated with overexpression of the gene product, indicating that the polypeptides are useful targets for therapeutic intervention in certain cancers such as colon, lung, breast and other cancers and diagnostic determination of the presence of those cancers" (emphasis added). Besides, Appellants have submitted ample evidence (discussed below) to show that, in general, if a gene is amplified in cancer, it is "more likely than not" likely that the encoded protein will also be expressed at an elevated level.

For support, Appellants presented the articles by Orntoft *et al.*, Hyman *et al.*, and Pollack *et al.* (made of record in Appellants' Response filed August 3, 2004), who collectively teach that in general, for most genes, DNA amplification increases mRNA expression. The results presented by Orntoft *et al.*, Hyman *et al.*, and Pollack *et al.* are based upon wide ranging analyses of a large number of tumor associated genes. Orntoft *et al.* studied transcript levels of 5600 genes in malignant bladder cancers, many of which were linked to the gain or loss of chromosomal material, and found that in general (18 of 23 cases) chromosomal areas with more than 2-fold gain of DNA showed a corresponding increase in mRNA transcripts. Hyman *et al.* compared DNA copy numbers and mRNA expression of over 12,000 genes in breast cancer tumors and cell lines, and found that there was evidence of a prominent global influence of copy number changes on gene expression levels. In Pollack *et al.*, the authors profiled DNA copy number alteration across 6,691 mapped human genes in 44 predominantly advanced primary breast tumors and 10 breast cancer cell lines, and found that on average, a 2-fold change in DNA copy number was associated with a corresponding 1.5-fold change in mRNA levels. In summary,



the evidence supports the Appellants' position that gene amplification is more likely than not predictive of increased mRNA and polypeptide levels.

Second, the Declaration of Dr. Paul Polakis (made of record in Appellants' Response filed August 3, 2004), principal investigator of the Tumor Antigen Project of Genentech, Inc., the assignee of the present application, explains that in the course of Dr. Polakis' research using microarray analysis, he and his co-workers identified approximately 200 gene transcripts that are present in human tumor cells at significantly higher levels than in corresponding normal human cells. Appellants submit that Dr. Polakis' Declaration was presented to support the position that there is a correlation between mRNA levels and polypeptide levels, the correlation between gene amplification and mRNA levels having already been established by the data shown in the Orntoft *et al.*, Hyman *et al.*, and Pollack *et al.* articles. Appellants further emphasize that the opinions expressed in the Polakis Declaration, including in the above quoted statement, are all based on factual findings. For instance, antibodies binding to about 30 of these tumor antigens were prepared, and mRNA and protein levels were compared. In approximately 80% of the cases, the researchers found that increases in the level of a particular mRNA correlated with changes in the level of protein expressed from that mRNA when human tumor cells are compared with their corresponding normal cells. Therefore, Dr. Polakis' research, which is referenced in his Declaration, shows that, in general, there is a correlation between increased mRNA and polypeptide levels.

Appellants further note that the sale of gene expression chips to measure mRNA levels is a highly successful business, with a company such as Affymetrix recording 168.3 million dollars in sales of their GeneChip® arrays in 2004. Clearly, the research community believe that the information obtained from these chips is useful (i.e., that it is more likely than not that the results are informative of protein levels).

Taken together, all of the submitted evidence supports the Appellants' position that, in the majority of amplified genes, increased gene amplification levels, more likely than not, predict increased mRNA and polypeptide levels, which clearly meets the utility standards described above. Hence, one of skill in the art would reasonably expect that, based on the gene amplification data of the PRO1097 gene, the PRO1097 polypeptide is concomitantly overexpressed in the lung or colon tumors studied as well.

Appellants further submit that, even if there were no correlation between gene amplification and increased mRNA/protein expression, (which Appellants expressly do not concede), a polypeptide encoded by an amplified gene in cancer would **still** have a specific, substantial, and credible utility as explained below. As the Declaration of Dr. Avi Ashkenazi (submitted with Appellants' Response filed August 3, 2004) explains:

"even when amplification of a cancer marker gene does not result in significant over-expression of the corresponding gene product, this very absence of gene product over-expression still provides significant information for cancer diagnosis and treatment."

Thus, even if over-expression of the gene product does not parallel gene amplification in certain tumor types, parallel monitoring of gene amplification and gene product over-expression enables more accurate tumor classification and hence better determination of suitable therapy. In addition, absence of over-expression is crucial information for the practicing clinician. If a gene is amplified in a tumor, but the corresponding gene product is not over-expressed, the clinician will decide not to treat a patient with agents that target that gene product. This not only saves money, but also has the benefit that the patient can avoid exposure to the side effects associated with such agents.

This utility is further supported by the teachings of the article by Hanna and Mornin. (Pathology Associates Medical Laboratories, August (1999), submitted with the Response filed August 3, 2004). The article teaches that the HER-2/neu gene has been shown to be amplified and/or over-expressed in 10%-30% of invasive breast cancers and in 40%-60% of intraductal breast carcinomas. Further, the article teaches that diagnosis of breast cancer includes testing both the amplification of the HER-2/neu gene (by FISH) as well as the over-expression of the HER-2/neu gene product (by IHC). Even when the protein is not over-expressed, the assay relying on both tests leads to a more accurate classification of the cancer and a more effective treatment of it.

The Examiner asserts that,

"Hanna *et al.* supports the instant rejection, in that Hanna *et al.* show that gene amplification does not reliably correlate with polypeptide overexpression, and thus the level of polypeptide expression must be tested empirically." (Page 8 of the Final Office Action mailed October 18, 2004).

Appellants respectfully point out that the Examiner appears to have misread Hanna *et al.* Hanna *et al.* clearly state that gene amplification (as measured by FISH) and polypeptide expression (as measured by immunohistochemistry, IHC) are well correlated ("in general, FISH and IHC results correlate well" (Hanna *et al.* p. 1, col. 2)). It is only a subset of tumors which show discordant results. Thus, Hanna *et al.* support Appellants' position rather well that it is more likely than not that gene amplification correlates with increased polypeptide expression. The Examiner appears to view such testing described in the Ashkenazi Declaration and the Hanna paper as experiments involving further characterization of the PRO1097 polypeptide itself. On the contrary, such testing is for the purpose of characterizing not the PRO1097 polypeptide, but the tumors in which the gene encoding PRO1097 is amplified. That is, such further testing or research is for the purpose of characterizing the tumors into medically relevant categories in which the gene encoding PRO1097 is/is not amplified, and such techniques were routine in the art of clinical oncology at the time of filing of the instant application, as evidenced by the teaching of Hanna and Mornin.

Thus, based on the asserted utility for PRO1097 in the diagnosis of selected lung or colon tumors, the reduction to practice of the instantly claimed protein sequence of SEQ ID NO: 349 in the present application (also see page 305), the disclosure of the step-by-step protocols for making chimeric PRO polypeptides, including those wherein the heterologous polypeptide is an epitope tag or an Fc region of an immunoglobulin in the specification (at page 374, lines 24 to page 375, line 9), the disclosure of a step-by-step protocol for making and expressing PRO1097 in appropriate host cells (in Examples 140-143 and page 376, line 12), the step-by-step protocol for the preparation, isolation and detection of monoclonal, polyclonal and other types of antibodies against the PRO1097 protein in the specification (at pages 390-395) and the disclosure of the gene amplification assay in Example 170, the skilled artisan would know exactly how to make and use the claimed polypeptide for the diagnosis of lung or colon cancers. Appellants submit that based on the detailed information presented in the specification and the advanced state of the art in oncology, the skilled artisan would have found such testing routine and not 'undue'.

Contrary to the Appellants assertion of utility, however, the Examiner alleges that the gene amplification results presented in Example 170 does not render the presently claimed

polypeptides patentably useful, and, finds the declaratory evidence presented in this case, for what Appellants consider legally inappropriate reasons, "non-persuasive".

Regarding the non-acceptance of the Polakis and Ashkenazi declarations by the Examiner, Appellants respectfully draw the Examiner's attention to case law that clearly establishes that in considering affidavit evidence, the Examiner must consider all of the evidence of record anew (*In re Rinehart*, 531 F.2d 1084, 189 USPQ 143 (C.C.P.A. 1976) and *In re Piasecki*, 745 F.2d 1015, 226 USPQ 881 (Fed. Cir. 1985)). "After evidence or argument is submitted by the applicant in response, patentability is determined on the totality of the record, by a preponderance of the evidence with due consideration to persuasiveness of argument" (*In re Alton*, 37 USPQ2d 1578 (Fed. Cir 1966) at 1584 quoting *In re Oetiker*, 977 F.2d 1443, 1445, 24 USPQ2d 1443, 1444 (Fed. Cir. 1992)). Furthermore, the Federal Court of Appeals held in *In re Alton*, "We are aware of no reason why opinion evidence relating to a fact issue should not be considered by an examiner" (*In re Alton, supra.*). Appellants further draw the Examiner's attention to the Utility Examination Guidelines (Part IIB, 66 Fed. Reg. 1098 (2001)) which states,

"Office personnel must accept an opinion from a qualified expert that is based upon relevant facts whose accuracy is not being questioned; it is improper to disregard the opinion solely because of a disagreement over the significance or meaning of the facts offered."

The statement in question from the Polakis Declaration that "it is my considered scientific opinion that for human genes, an increased level of mRNA in a tumor cell relative to a normal cell typically correlates to a similar increase in abundance of the encoded protein in the tumor cell relative to the normal cell" is based on his own experimental findings, which is clearly set forth in the Declaration. Further, the teachings of Ashkenazi were supported by the Her-2/neu gene example in Hanna and Mornin. Accordingly, the fact-based conclusions of Dr. Polakis and Dr. Ashkenazi would be considered reasonable and accurate by one skilled in the art. Thus, barring evidence to the contrary, Appellants maintain that the fold amplification disclosed for the PRO1097 gene is significant and forms the basis for the utility claimed for the PRO1097 polypeptide herein.

Therefore, since the instantly claimed invention is supported by either a credible, specific and substantial asserted utility or a well-established utility, and since the present specification

clearly teaches one skilled in the art "how to make and use" the claimed invention without undue experimentation, Appellants respectfully request reconsideration and reversal of this outstanding rejections under 35 U.S.C. §101 and §112, First Paragraph to Claims 119-126 and 129-131.

**ISSUE 3: Claims 119-126 satisfy the written description requirement of 35 U.S.C. §112,**

**First Paragraph**

Claims 119-126 stand rejected under 35 U.S.C. §112, first paragraph as allegedly containing "subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention." In particular, the Examiner has asserted that "Applicants have not described or shown possession of all polypeptides 80-99% homologous to SEQ ID NO:349, that still retain the function of SEQ ID NO: 349. Nor have Applicants described a representative number of species that have 80-99% homology to SEQ ID NO: 349, such that it is clear that they were in possession of a genus of polypeptides functionally similar to SEQ ID NO: 349" (Page 9 of the Final Office Action mailed October 20, 2004).

Appellants respectfully disagree.

**A. The Legal Test for Written Description**

The well-established test for sufficiency of support under the written description requirement of 35 U.S.C. §112, first paragraph is "whether the disclosure of the application as originally filed reasonably conveys to the artisan that the inventor had possession at that time of the later claimed subject matter, rather than the presence or absence of literal support in the specification for the claim language."<sup>18,19</sup> The adequacy of written description support is a factual issue and is to be determined on a case-by-case basis.<sup>20</sup> The factual determination in a

---

<sup>18</sup> *In re Kaslow*, 707 F.2d 1366, 1374, 212 USPQ 1089, 1096 (Fed. Cir. 1983).

<sup>19</sup> *See also Vas-Cath, Inc. v. Mahurkar*, 935 F.2d at 1563, 19 USPQ2d at 1116 (Fed. Cir. 1991).

<sup>20</sup> *See e.g., Vas-Cath*, 935 F.2d at 1563; 19 USPQ2d at 1116.

written description analysis depends on the nature of the invention and the amount of knowledge imparted to those skilled in the art by the disclosure.<sup>21, 22</sup>

In *Environmental Designs, Ltd. v. Union Oil Co.*,<sup>23</sup> the Federal Circuit held, "Factors that may be considered in determining level of ordinary skill in the art include (1) the educational level of the inventor; (2) type of problems encountered in the art; (3) prior art solutions to those problems; (4) rapidity with which innovations are made; (5) sophistication of the technology; and (6) educational level of active workers in the field."<sup>24</sup> Further, the "hypothetical 'person having ordinary skill in the art' to which the claimed subject matter pertains would, of necessity have the capability of understanding the scientific and engineering principles applicable to the pertinent art."<sup>25, 26</sup>

**B. The Disclosure Provides Sufficient Written Description for the Claimed Invention**

Appellants respectfully submit that the instant specification evidences the actual reduction to practice of the amino acid sequence of SEQ ID NO: 349. Appellants also submit that the specification provides ample written support for determining percent sequence identity between two amino acid sequences (See pages 306-308, line 14 onwards). In fact, the specification teaches specific parameters to be associated with the term "percent identity" as applied to the present invention. The specification further provides detailed guidance as to changes that may be made to a PRO polypeptide without adversely affecting its activity (page

---

<sup>21</sup> *Union Oil v. Atlantic Richfield Co.*, 208 F.2d 989, 996 (Fed. Cir. 2000).

<sup>22</sup> *See also* M.P.E.P. §2163 II(A).

<sup>23</sup> 713 F.2d 693, 696, 218 USPQ 865, 868 (Fed. Cir. 1983), *cert. denied*, 464 U.S. 1043 (1984).

<sup>24</sup> *See also* M.P.E.P. §2141.03.

<sup>25</sup> *Ex parte Hiyamizu*, 10 USPQ2d 1393, 1394 (Bd. Pat. App. & Inter. 1988) (emphasis added).

<sup>26</sup> *See also* M.P.E.P. §2141.03.

372, line 36 to page 373, line 17). This guidance includes a listing of exemplary and preferred substitutions for each of the twenty naturally occurring amino acids (Table 6, page 372). Accordingly, one of skill in the art could identify whether the variant PRO1097 sequence falls within the parameters of the claimed invention. Once such an amino acid sequence was identified, the specification sets forth methods for making the amino acid sequences (see page 376, line 9) and methods of preparing the PRO polypeptides (see Example 140-143).

Currently pending Claims 119-126 recite the functional limitation that the nucleic acid encoding the claimed polypeptides are amplified in lung or colon tumors. Appellants further submit that the specification provides ample written support for detecting and quantifying amplification of such nucleic acids in several tumors and/or cell lines as described in Example 170. Example 170 of the present application provides step-by-step guidelines and protocols for the gene amplification assay. By following this disclosure, one skilled in the art would know that it is easy to test whether a gene encoding a variant PRO1097 protein is amplified in lung or colon tumors by the methods set forth in Example 170.

Thus, the genus of polypeptides with at least 80% sequence identity to SEQ ID NO: 349, which possess the functional property of having a nucleic acid which is amplified in lung or colon tumor would meet the requirement of 35 U.S.C. §112, first paragraph, as providing adequate written description. Accordingly, one skilled in the art would have known that Appellants had knowledge and possessed the claimed polypeptides with 80-99% sequence identity to SEQ ID NO: 349 whose encoding nucleic acids were amplified in lung or colon tumors. The recited property of amplification of the encoding gene adds to the characterization of the claimed polypeptide sequences in a manner that one of skill in the art could readily assess and understand.

As discussed above, Appellants have recited structural features, namely, 80% sequence identity to SEQ ID NO: 349, which are common to the genus. Appellants have also provided guidance as to how to make the recited variants of SEQ ID NO: 349, including listings of exemplary and preferred sequence substitutions. The genus of claimed polypeptides is further defined by having a specific functional activity for the encoding nucleic acids. Accordingly, a description of the claimed genus has been achieved.

For the above reasons, the specification provides adequate written description for polypeptides having at least 80% identity to SEQ ID NO: 349 wherein the nucleic acid encoding

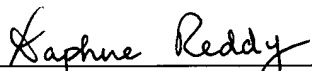
the polypeptide is amplified in lung or colon tumor. Accordingly, Appellants respectfully request reconsideration and reversal of the written description rejection of Claims 119-126 under 35 U.S.C. §112, first paragraph.

### **CONCLUSION**

For the reasons given above, Appellants submit that present specification clearly describes, details and provides a patentable utility for the claimed invention. Moreover, it is respectfully submitted that based upon this disclosed patentable utility, the present specification clearly teaches "how to use" the presently claimed polypeptide. As such, Appellants respectfully request reconsideration and reversal of the outstanding rejection of claims 119-126 and 129-131.

Respectfully submitted,

Date: October 28, 2005

  
\_\_\_\_\_  
Daphne Reddy (Reg. No. 53,507)

**HELLER EHRMAN LLP**  
275 Middlefield Road  
Menlo Park, California 94025-3506  
Telephone: (650) 324-7000  
Facsimile: (650) 324-0638



## **IX. CLAIMS APPENDIX**

### **Claims on Appeal**

119. An isolated polypeptide having at least 80% amino acid sequence identity to:
- (a) the amino acid sequence of the polypeptide of SEQ ID NO:349;
  - (b) the amino acid sequence of the polypeptide of SEQ ID NO:349, lacking its associated signal peptide; or
  - (c) the amino acid sequence of the polypeptide encoded by the full-length coding sequence of the cDNA deposited under ATCC accession number 203044;
- wherein, the nucleic acid encoding said polypeptide is amplified in lung or colon cancer.
120. An isolated polypeptide having at least 85% amino acid sequence identity to:
- (a) the amino acid sequence of the polypeptide of SEQ ID NO:349;
  - (b) the amino acid sequence of the polypeptide of SEQ ID NO:349, lacking its associated signal peptide; or
  - (c) the amino acid sequence of the polypeptide encoded by the full-length coding sequence of the cDNA deposited under ATCC accession number 203044;
- wherein, the nucleic acid encoding said polypeptide is amplified in lung or colon cancer.
121. An isolated polypeptide having at least 90% amino acid sequence identity to:
- (a) the amino acid sequence of the polypeptide of SEQ ID NO:349;
  - (b) the amino acid sequence of the polypeptide of SEQ ID NO:349, lacking its associated signal peptide; or
  - (c) the amino acid sequence of the polypeptide encoded by the full-length coding sequence of the cDNA deposited under ATCC accession number 203044;
- wherein, the nucleic acid encoding said polypeptide is amplified in lung or colon cancer.
122. An isolated polypeptide having at least 95% amino acid sequence identity to:
- (a) the amino acid sequence of the polypeptide of SEQ ID NO:349;
  - (b) the amino acid sequence of the polypeptide of SEQ ID NO:349, lacking its associated signal peptide; or

(c) the amino acid sequence of the polypeptide encoded by the full-length coding sequence of the cDNA deposited under ATCC accession number 203044;

wherein, the nucleic acid encoding said polypeptide is amplified in lung or colon cancer.

123. An isolated polypeptide having at least 99% amino acid sequence identity to:

(a) the amino acid sequence of the polypeptide of SEQ ID NO:349;

(b) the amino acid sequence of the polypeptide of SEQ ID NO:349, lacking its associated signal peptide; or

(c) the amino acid sequence of the polypeptide encoded by the full-length coding sequence of the cDNA deposited under ATCC accession number 203044;

wherein, the nucleic acid encoding said polypeptide is amplified in lung or colon cancer.

124. An isolated polypeptide comprising:

(a) the amino acid sequence of the polypeptide of SEQ ID NO: 349;

(b) the amino acid sequence of the polypeptide of SEQ ID NO: 349, lacking its associated signal peptide;

(c) the amino acid sequence of the polypeptide encoded by the full-length coding sequence of the cDNA deposited under ATCC accession number 203044;

wherein, the nucleic acid encoding said polypeptide is amplified in lung or colon cancer.

125. The isolated polypeptide of Claim 124 comprising the amino acid sequence of the polypeptide of SEQ ID NO: 349.

126. The isolated polypeptide of Claim 124 comprising the amino acid sequence of the polypeptide of SEQ ID NO: 349, lacking its associated signal peptide.

129. The isolated polypeptide of Claim 124 comprising the amino acid sequence of the polypeptide encoded by the full-length coding sequence of the cDNA deposited under ATCC accession number 203044.

130. A chimeric polypeptide comprising a polypeptide according to Claim 124 fused to a heterologous polypeptide.

131. The chimeric polypeptide of Claim 130, wherein said heterologous polypeptide is an epitope tag or an Fc region of an immunoglobulin.

## **X. EVIDENCE APPENDIX**

1. Declaration of Paul Polakis, Ph.D. under 35 C.F.R. 1.132.
2. Declaration of Avi Ashkenazi, Ph.D. under 35 C.F.R. 1.132.
3. Declaration of Audrey Goddard, Ph.D. under 35 C.F.R. 1.132.
4. Orntoft *et al.*, 2002, Mol. and Cell. Proteomics, Vol.1, pages 37-45.
5. Hyman *et al.*, Cancer Res., 2002, Vol. 62, pages 6240-45.
6. Pollack *et al.*, PNAS, 2002, Vol. 99, pages 12963-12968.
7. Hanna and Mornin, 1999, Pathology Associates Medical Laboratories.
8. Hittelman *et al.* 2001, Ann. N.Y. Acad. Sci. 952: 1-12.
9. Crowell *et al.* 1996, Cancer Epidemiol., 5: 631-37.
10. Skolnick *et al.*, 2000, Trends in Biotech., 18:34-39.
11. Bork *et al.*, 2000, Genome Res. 10: 398-400.
12. Doerks *et al.*, 1998, Trends in Genetics, 14: 248-250.
13. Hesselgesser *et al.*, 1997, Meth. in Enzymol., 287: 59-69.
14. Blease *et al.*, 2000, Resp. Res., 1(1): 54-61.
15. Smith *et al.*, 1997, Nat. Biotechnol., 15: 1222-23.
16. Brenner *et al.*, 1999, Trends in Genetics, 15: 132-133.
17. Pennica *et al.*, Proc. Nat. Acad. Sci., 1998, Vol. 95, pages 141097-722.
18. Haynes *et al.*, Electrophoresis, 1998, Vol. 19, pages 1862-71.
19. Hu *et al.*, J. Proteome Res., 2003, Vol. 2, pages 405-412.

Items 1, 2, 4-7 were submitted with Appellants' Response filed August 3, 2004, and were considered by the Examiner as indicated in the Final Office action mailed October 18, 2004.

Item 3 is hereby submitted with the Appellants' brief. As indicated above, this declaration was not presented earlier because the issue whether the "fold increase" in the gene amplification assay was "significant" was not raised earlier. However, Appellants believe that presentation of the Goddard Declaration as evidence that the "increase in DNA" in the gene amplification assay is significant is necessary in this case and presents the case in better form for appeal. Its consideration is respectfully requested.

Items 8-16 were made of record by the Examiner in the Office Action mailed May 3, 2004.

Items 17-19 were made of record by the Examiner in the Final Office Action mailed October 18, 2004.

**XI. RELATED PROCEEDINGS APPENDIX**

None.



## DECLARATION OF PAUL POLAKIS, Ph.D.

Paul Polakis, Ph.D., declare and say as follows:

1. I was awarded a Ph.D. by the Department of Biochemistry of the Michigan State University in 1984. My scientific Curriculum Vitae is attached to and forms part of this Declaration (Exhibit A).
2. I am currently employed by Genentech, Inc. where my job title is Staff Scientist. Since joining Genentech in 1999, one of my primary responsibilities has been leading Genentech's Tumor Antigen Project, which is a large research project with a primary focus on identifying tumor cell markers that find use as targets for both the diagnosis and treatment of cancer in humans.
3. As part of the Tumor Antigen Project, my laboratory has been analyzing differential expression of various genes in tumor cells relative to normal cells. The purpose of this research is to identify proteins that are abundantly expressed on certain tumor cells and that are either (i) not expressed, or (ii) expressed at lower levels, on corresponding normal cells. We call such differentially expressed proteins "tumor antigen proteins". When such a tumor antigen protein is identified, one can produce an antibody that recognizes and binds to that protein. Such an antibody finds use in the diagnosis of human cancer and may ultimately serve as an effective therapeutic in the treatment of human cancer.
4. In the course of the research conducted by Genentech's Tumor Antigen Project, we have employed a variety of scientific techniques for detecting and studying differential gene expression in human tumor cells relative to normal cells, at genomic DNA, mRNA and protein levels. An important example of one such technique is the well known and widely used technique of microarray analysis which has proven to be extremely useful for the identification of mRNA molecules that are differentially expressed in one tissue or cell type relative to another. In the course of our research using microarray analysis, we have identified approximately 200 gene transcripts that are present in human tumor cells at significantly higher levels than in corresponding normal human cells. To date, we have generated antibodies that bind to about 30 of the tumor antigen proteins expressed from these differentially expressed gene transcripts and have used these antibodies to quantitatively determine the level of production of these tumor antigen proteins in both human cancer cells and corresponding normal cells. We have then compared the levels of mRNA and protein in both the tumor and normal cells analyzed.
5. From the mRNA and protein expression analyses described in paragraph 4 above, we have observed that there is a strong correlation between changes in the level of mRNA present in any particular cell type and the level of protein

expressed from that mRNA in that cell type. In approximately 80% of our observations we have found that increases in the level of a particular mRNA correlates with changes in the level of protein expressed from that mRNA when human tumor cells are compared with their corresponding normal cells.

6. Based upon my own experience accumulated in more than 20 years of research, including the data discussed in paragraphs 4 and 5 above and my knowledge of the relevant scientific literature, it is my considered scientific opinion that for human genes, an increased level of mRNA in a tumor cell relative to a normal cell typically correlates to a similar increase in abundance of the encoded protein in the tumor cell relative to the normal cell. In fact, it remains a central dogma in molecular biology that increased mRNA levels are predictive of corresponding increased levels of the encoded protein. While there have been published reports of genes for which such a correlation does not exist, it is my opinion that such reports are exceptions to the commonly understood general rule that increased mRNA levels are predictive of corresponding increased levels of the encoded protein.

7. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information or belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful statements may jeopardize the validity of the application or any patent issued thereon.

Dated: 5/07/04

By: Paul Polakis

Paul Polakis, Ph.D.



## CURRICULUM VITAE

PAUL G. POLAKIS  
Staff Scientist  
Genentech, Inc  
1 DNA Way, MS#40  
S. San Francisco, CA 94080

### EDUCATION:

Ph.D., Biochemistry, Department of Biochemistry,  
Michigan State University (1984)

B.S., Biology. College of Natural Science, Michigan State University (1977)

### PROFESSIONAL EXPERIENCE:

2002-present	Staff Scientist, Genentech, Inc S. San Francisco, CA
1999- 2002	Senior Scientist, Genentech, Inc., S. San Francisco, CA
1997 -1999	Research Director Onyx Pharmaceuticals, Richmond, CA
1992- 1996	Senior Scientist, Project Leader, Onyx Pharmaceuticals, Richmond, CA
1991-1992	Senior Scientist, Chiron Corporation, Emeryville, CA.
1989-1991	Scientist, Cetus Corporation, Emeryville CA.
1987-1989	Postdoctoral Research Associate, Genentech, Inc., South San Francisco, CA.
1985-1987	Postdoctoral Research Associate, Department of Medicine, Duke University Medical Center, Durham, NC

1984-1985

Assistant Professor, Department of Chemistry,  
Oberlin College, Oberlin, Ohio

1980-1984

Graduate Research Assistant, Department of  
Biochemistry, Michigan State University  
East Lansing, Michigan

### **PUBLICATIONS:**

1. **Polakis, P. G.** and Wilson, J. E. 1982 Purification of a Highly Bindable Rat Brain Hexokinase by High Performance Liquid Chromatography. **Biochem. Biophys. Res. Commun.** 107, 937-943.
2. **Polakis, P.G.** and Wilson, J. E. 1984 Proteolytic Dissection of Rat Brain Hexokinase: Determination of the Cleavage Pattern during Limited Digestion with Trypsin. **Arch. Biochem. Biophys.** 234, 341-352.
3. **Polakis, P. G.** and Wilson, J. E. 1985 An Intact Hydrophobic N-Terminal Sequence is Required for the Binding Rat Brain Hexokinase to Mitochondria. **Arch. Biochem. Biophys.** 236, 328-337.
4. Uhing, R.J., **Polakis, P.G.** and Snyderman, R. 1987 Isolation of GTP-binding Proteins from Myeloid HL60 Cells. **J. Biol. Chem.** 262, 15575-15579.
5. **Polakis, P.G.**, Uhing, R.J. and Snyderman, R. 1988 The Formylpeptide Chemoattractant Receptor Copurifies with a GTP-binding Protein Containing a Distinct 40 kDa Pertussis Toxin Substrate. **J. Biol. Chem.** 263, 4969-4979.
6. Uhing, R. J., Dillon, S., **Polakis, P. G.**, Truett, A. P. and Snyderman, R. 1988 Chemoattractant Receptors and Signal Transduction Processes in Cellular and Molecular Aspects of Inflammation ( Poste, G. and Crooke, S. T. eds.) pp 335-379.
7. **Polakis, P.G.**, Evans, T. and Snyderman 1989 Multiple Chromatographic Forms of the Formylpeptide Chemoattractant Receptor and their Relationship to GTP-binding Proteins. **Biochem. Biophys. Res. Commun.** 161, 276-283.
8. **Polakis, P. G.**, Snyderman, R. and Evans, T. 1989 Characterization of G25K, a GTP-binding Protein Containing a Novel Putative Nucleotide Binding Domain. **Biochem. Biophys. Res. Commun.** 160, 25-32.
9. **Polakis, P.**, Weber, R.F., Nevins, B., Didsbury, J. Evans, T. and Snyderman, R. 1989 Identification of the ral and rac1 Gene Products, Low Molecular Mass GTP-binding Proteins from Human Platelets. **J. Biol. Chem.** 264, 16383-16389.
10. Snyderman, R., Perianin, A., Evans, T., **Polakis, P.** and Didsbury, J. 1989 G Proteins and Neutrophil Function. In ADP-Ribosylating Toxins and G Proteins: Insights into Signal Transduction. ( J. Moss and M. Vaughn, eds.) Amer. Soc. Microbiol. pp. 295-323.

11. Hart, M.J., **Polakis, P.G.**, Evans, T. and Cerrione, R.A. 1990 The Identification and Characterization of an Epidermal Growth Factor-Stimulated Phosphorylation of a Specific Low Molecular Mass GTP-binding Protein in a Reconstituted Phospholipid Vesicle System. **J. Biol. Chem.** 265, 5990-6001.
12. Yatani, A., Okabe, K., **Polakis, P.** Halenbeck, R. McCormick, F. and Brown, A. M. 1990 ras p21 and GAP Inhibit Coupling of Muscarinic Receptors to Atrial K<sup>+</sup> Channels. **Cell.** 61, 769-776.
13. Munemitsu, S., Innis, M.A., Clark, R., McCormick, F., Ullrich, A. and **Polakis, P.G.** 1990 Molecular Cloning and Expression of a G25K cDNA, the Human Homolog of the Yeast Cell Cycle Gene CDC42. **Mol. Cell. Biol.** 10, 5977-5982.
14. **Polakis, P.G.** Rubinfeld, B. Evans, T. and McCormick, F. 1991 Purification of Plasma Membrane-Associated GTPase Activating Protein (GAP) Specific for rap-1/krev-1 from HL60 Cells. **Proc. Natl. Acad. Sci. USA** 88, 239-243.
15. Moran, M. F., **Polakis, P.**, McCormick, F., Pawson, T. and Ellis, C. 1991 Protein Tyrosine Kinases Regulate the Phosphorylation, Protein Interactions, Subcellular Distribution, and Activity of p21ras GTPase Activating Protein. **Mol. Cell. Biol.** 11, 1804-1812
16. Rubinfeld, B., Wong, G., Bekesi, E. Wood, A. McCormick, F. and **Polakis, P. G.** 1991 A Synthetic Peptide Corresponding to a Sequence in the GTPase Activating Protein Inhibits p21<sup>ras</sup> Stimulation and Promotes Guanine Nucleotide Exchange. **Internatl. J. Peptide and Prot. Res.** 38, 47-53.
17. Rubinfeld, B., Munemitsu, S., Clark, R., Conroy, L., Watt, K., Crosier, W., McCormick, F., and **Polakis, P.** 1991 Molecular Cloning of a GTPase Activating Protein Specific for the Krev-1 Protein p21<sup>rap1</sup>. **Cell** 65, 1033-1042.
18. Zhang, K. Papageorge, A., G., Martin, P., Vass, W. C., Olah, Z., **Polakis, P.**, McCormick, F. and Lowy, D. R. 1991 Heterogenous Amino Acids in RAS and Rap1A Specifying Sensitivity to GAP Proteins. **Science** 254, 1630-1634.
19. Martin, G., Yatani, A., Clark, R., **Polakis, P.**, Brown, A. M. and McCormick, F. 1992 GAP Domains Responsible for p21<sup>ras</sup>-dependent Inhibition of Muscarinic Atrial K<sup>+</sup> Channel Currents. **Science** 255, 192-194.
20. McCormick, F., Martin, G. A., Clark, R., Bollag, G. and **Polakis, P.** 1992 Regulation of p21ras by GTPase Activating Proteins. Cold Spring Harbor **Symposia on Quantitative Biology**. Vol. 56, 237-241.
21. Pronk, G. B., **Polakis, P.**, Wong, G., deVries-Smits, A. M., Bos J. L. and McCormick, F. 1992 p60<sup>v-src</sup> Can Associate with and Phosphorylate the p21<sup>ras</sup> GTPase Activating Protein. **Oncogene** 7,389-394.
22. **Polakis P.** and McCormick, F. 1992 Interactions Between p21<sup>ras</sup> Proteins and Their GTPase Activating Proteins. In **Cancer Surveys** ( Franks, L. M., ed.) 12, 25-42.

23. Wong, G., Muller, O., Clark, R., Conroy, L., Moran, M., **Polakis, P.** and McCormick, F. 1992 Molecular cloning and nucleic acid binding properties of the GAP-associated tyrosine phosphoprotein p62. **Cell** 69, 551-558.
24. **Polakis, P.**, Rubinfeld, B. and McCormick, F. 1992 Phosphorylation of rap1GAP in vivo and by cAMP-dependent Kinase and the Cell Cycle p34<sup>cdc2</sup> Kinase in vitro. **J. Biol. Chem.** 267, 10780-10785.
25. McCabe, P.C., Haubrauck, H., **Polakis, P.**, McCormick, F., and Innis, M. A. 1992 Functional Interactions Between p21<sup>rap1A</sup> and Components of the Budding pathway of *Saccharomyces cerevisiae*. **Mol. Cell. Biol.** 12, 4084-4092.
26. Rubinfeld, B., Crosier, W.J., Albert, I., Conroy, L., Clark, R., McCormick, F. and **Polakis, P.** 1992 Localization of the rap1GAP Catalytic Domain and Sites of Phosphorylation by Mutational Analysis. **Mol. Cell. Biol.** 12, 4634-4642.
27. Ando, S., Kaibuchi, K., Sasaki, K., Hiraoka, T., Nishiyama, T., Mizuno, T., Asada, M., Nunoi, H., Matsuda, I., Matsuura, Y., **Polakis, P.**, McCormick, F. and Takai, Y. 1992 Post-translational processing of rac p21s is important both for their interaction with the GDP/GTP exchange proteins and for their activation of NADPH oxidase. **J. Biol. Chem.** 267, 25709-25713.
28. Janoueix-Lerosey, I., **Polakis, P.**, Tavitian, A. and deGunzberg, J. 1992 Regulation of the GTPase activity of the ras-related rap2 protein. **Biochem. Biophys. Res. Commun.** 189, 455-464.
29. **Polakis, P.** 1993 GAPs Specific for the rap1/Krev-1 Protein. in GTP-binding Proteins: the ras-superfamily. ( J.C. LaCale and F. McCormick, eds.) 445-452.
30. **Polakis, P.** and McCormick, F. 1993 Structural requirements for the interaction of p21<sup>ras</sup> with GAP, exchange factors, and its biological effector target. **J. Biol Chem.** 268, 9157-9160.
31. Rubinfeld, B., Souza, B. Albert, I., Muller, O., Chamberlain, S., Masiarz, F., Munemitsu, S. and **Polakis, P.** 1993 Association of the APC gene product with beta- catenin. **Science** 262, 1731-1734.
32. Weiss, J., Rubinfeld, B., **Polakis, P.**, McCormick, F. Cavenee, W. A. and Arden, K. 1993 The gene for human rap1-GTPase activating protein (rap1GAP) maps to chromosome 1p35-1p36.1. **Cytogenet. Cell Genet.** 66, 18-21.
33. Sato, K. Y., **Polakis, P.**, Haubruck, H., Fasching, C. L., McCormick, F. and Stanbridge, E. J. 1994 Analysis of the tumor suppressor activity of the K-rev gene in human tumor cell lines. **Cancer Res.** 54, 552-559.
34. Janoueix-Lerosey, I., Fontenay, M., Tobelem, G., Tavitian, A., **Polakis, P.** and DeGunzburg, J. 1994 Phosphorylation of rap1GAP during the cell cycle. **Biochem. Biophys. Res. Commun.** 202, 967-975
35. Munemitsu, S., Souza, B., Mueller, O., Albert, I., Rubinfeld, B., and **Polakis, P.** 1994 The APC gene product associates with microtubules in vivo and affects their assembly in vitro. **Cancer Res.** 54, 3676-3681.

36. Rubinfeld, B. and Polakis, P. 1995 Purification of baculovirus produced rap1GAP. **Methods Enz.** 255,31
37. Polakis, P. 1995 Mutations in the APC gene and their implications for protein structure and function. **Current Opinions in Genetics and Development** 5, 66-71
38. Rubinfeld, B., Souza, B., Albert, I., Munemitsu, S. and Polakis P. 1995 The APC protein and E-cadherin form similar but independent complexes with  $\alpha$ -catenin,  $\beta$ -catenin and Plakoglobin. **J. Biol. Chem.** 270, 5549-5555
39. Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B., and Polakis, P. 1995 Regulation of intracellular  $\beta$ -catenin levels by the APC tumor suppressor gene. **Proc. Natl. Acad. Sci.** 92, 3046-3050.
40. Lock, P., Fumagalli, S., Polakis, P. McCormick, F. and Courtneidge, S. A. 1996 The human p62 cDNA encodes Sam68 and not the rasGAP-associated p62 protein. **Cell** 84, 23-24.
41. Papkoff, J., Rubinfeld, B., Schryver, B. and Polakis, P. 1996 Wnt-1 regulates free pools of catenins and stabilizes APC-catenin complexes. **Mol. Cell. Biol.** 16, 2128-2134.
42. Rubinfeld, B., Albert, I., Porfiri, E., Fiol, C., Munemitsu, S. and Polakis, P. 1996 Binding of GSK3 $\beta$  to the APC- $\beta$ -catenin complex and regulation of complex assembly. **Science** 272, 1023-1026.
43. Munemitsu, S., Albert, I., Rubinfeld, B. and Polakis, P. 1996 Deletion of amino-terminal structure stabilizes  $\beta$ -catenin in vivo and promotes the hyperphosphorylation of the APC tumor suppressor protein. **Mol. Cell. Biol.** 16, 4088-4094.
44. Hart, M. J., Callow, M. G., Sousa, B. and Polakis P. 1996 IQGAP1, a calmodulin binding protein with a rasGAP related domain, is a potential effector for cdc42Hs. **EMBO J.** 15, 2997-3005.
45. Nathke, I. S., Adams, C. L., Polakis, P., Sellin, J. and Nelson, W. J. 1996 The adenomatous polyposis coli (APC) tumor suppressor protein is localized to plasma membrane sites involved in active epithelial cell migration. **J. Cell. Biol.** 134, 165-180.
46. Hart, M. J., Sharma, S., elMasry, N., Qui, R-G., McCabe, P., Polakis, P. and Bollag, G. 1996 Identification of a novel guanine nucleotide exchange factor for the rho GTPase. **J. Biol. Chem.** 271, 25452.
47. Thomas JE, Smith M, Rubinfeld B, Gutowski M, Beckmann RP, and Polakis P. 1996 Subcellular localization and analysis of apparent 180-kDa and 220-kDa proteins of the breast cancer susceptibility gene, BRCA1. **J. Biol. Chem.** 1996 271, 28630-28635
48. Hayashi, S., Rubinfeld, B., Souza, B., Polakis, P., Wieschaus, E., and Levine, A. 1997 A Drosophila homolog of the tumor suppressor adenomatous polyposis coli

down-regulates  $\beta$ -catenin but its zygotic expression is not essential for the regulation of armadillo. **Proc. Natl. Acad. Sci.** 94, 242-247.

49. Vleminckx, K., Rubinfeld, B., **Polakis, P.** and Gumbiner, B. 1997 The APC tumor suppressor protein induces a new axis in *Xenopus* embryos. **J. Cell. Biol.** 136, 411-420.

50. Rubinfeld, B., Robbins, P., El-Gamil, M., Albert, I., Porfiri, P. and **Polakis, P.** 1997 Stabilization of  $\beta$ -catenin by genetic defects in melanoma cell lines. **Science** 275, 1790-1792.

51. **Polakis, P.** The adenomatous polyposis coli (APC) tumor suppressor. 1997 **Biochem. Biophys. Acta**, 1332, F127-F147.

52. Rubinfeld, B., Albert, I., Porfiri, E., Munemitsu, S., and **Polakis, P.** 1997 Loss of  $\beta$ -catenin regulation by the APC tumor suppressor protein correlates with loss of structure due to common somatic mutations of the gene. **Cancer Res.** 57, 4624-4630.

53. Porfiri, E., Rubinfeld, B., Albert, I., Hovanes, K., Waterman, M., and **Polakis, P.** 1997 Induction of a  $\beta$ -catenin-LEF-1 complex by wnt-1 and transforming mutants of  $\beta$ -catenin. **Oncogene** 15, 2833-2839.

54. Thomas JE, Smith M, Tonkinson JL, Rubinfeld B, and **Polakis P.**, 1997 Induction of phosphorylation on BRCA1 during the cell cycle and after DNA damage. **Cell Growth Differ.** 8, 801-809.

55. Hart, M., de los Santos, R., Albert, I., Rubinfeld, B., and **Polakis P.**, 1998 Down regulation of  $\beta$ -catenin by human Axin and its association with the adenomatous polyposis coli (APC) tumor suppressor,  $\beta$ -catenin and glycogen synthase kinase 3 $\beta$ . **Current Biology** 8, 573-581.

56. **Polakis, P.** 1998 The oncogenic activation of  $\beta$ -catenin. **Current Opinions in Genetics and Development** 9, 15-21

57. Matt Hart, Jean-Paul Concordet, Irina Lassot, Iris Albert, Rico del los Santos, Herve Durand, Christine Perret, Bonnee Rubinfeld, Florence Margottin, Richard Benarous and **Paul Polakis.** 1999 The F-box protein  $\beta$ -TrCP associates with phosphorylated  $\beta$ -catenin and regulates its activity in the cell. **Current Biology** 9, 207-10.

58. Howard C. Crawford, Barbara M. Fingleton, Bonnee Rubinfeld, **Paul Polakis** and Lynn M. Matrisian 1999 The metalloproteinase matrilysin is a target of  $\beta$ -catenin transactivation in intestinal tumours. **Oncogene** 18, 2883-91.

59. Meng J, Glick JL, **Polakis P.**, Casey PJ. 1999 Functional interaction between Galpha(z) and Rap1GAP suggests a novel form of cellular cross-talk. **J Biol Chem.** 17, 36663-9

60. Vijayasurian Easwaran, Virginia Song, **Paul Polakis** and Steve Byers 1999 The ubiquitin-proteosome pathway and serine kinase activity modulate APC mediated regulation of  $\beta$ -catenin-LEF signaling. **J. Biol. Chem.** 274(23):16641-5.
- 61 **Polakis P**, Hart M and Rubinfeld B. 1999 Defects in the regulation of beta-catenin in colorectal cancer. **Adv Exp Med Biol.** 470, 23-32
- 62 Shen Z, Batzer A, Koehler JA, **Polakis P**, Schlessinger J, Lydon NB, Moran MF. 1999 Evidence for SH3 domain directed binding and phosphorylation of Sam68 by Src. **Oncogene.** 18, 4647-53
64. Thomas GM, Frame S, Goedert M, Nathke I, **Polakis P**, Cohen P. 1999 A GSK3- binding peptide from FRAT1 selectively inhibits the GSK3-catalysed phosphorylation of axin and beta-catenin. **FEBS Lett.** 458, 247-51.
65. Peifer M, **Polakis P**. 2000 Wnt signaling in oncogenesis and embryogenesis--a look outside the nucleus. **Science** 287,1606-9.
66. **Polakis P**. 2000 Wnt signaling and cancer. **Genes Dev**;14, 1837-1851.
67. Spink KE, **Polakis P**, Weis WI 2000 Structural basis of the Axin-adenomatous polyposis coli interaction. **EMBO J** 19, 2270-2279.
68. Szeto, W., Jiang, W., Tice, D.A., Rubinfeld, B., Hollingshead, P.G., Fong, S.E., Dugger, D.L., Pham, T., Yansura, D.E., Wong, T.A., Grimaldi, J.C., Corpuz, R.T., Singh J.S., Frantz, G.D., Devaux, B., Crowley, C.W., Schwall, R.H., Eberhard, D.A., Rastelli, L., **Polakis, P.** and Pennica, D. 2001 Overexpression of the Retinoic Acid-Responsive Gene Stra6 in Human Cancers and its Synergistic Induction by Wnt-1 and Retinoic Acid. **Cancer Res** 61, 4197-4204.
69. Rubinfeld B, Tice DA, **Polakis P**. 2001 Axin dependent phosphorylation of the adenomatous polyposis coli protein mediated by casein kinase 1 epsilon. **J Biol Chem** 276, 39037-39045.
70. **Polakis P**. 2001 More than one way to skin a catenin. **Cell** 2001 105, 563-566.
71. Tice DA, Soloviev I, **Polakis P**. 2002 Activation of the Wnt Pathway Interferes with Serum Response Element-driven Transcription of Immediate Early Genes. **J Biol. Chem.** 277, 6118-6123.
72. Tice DA, Szeto W, Soloviev I, Rubinfeld B, Fong SE, Dugger DL, Winer J,

Williams PM, Wieand D, Smith V, Schwall RH, Pennica D, **Polakis P**. 2002 Synergistic activation of tumor antigens by wnt-1 signaling and retinoic acid revealed by gene expression profiling. **J Biol Chem**. 277,14329-14335.

73. **Polakis, P**. 2002 Casein kinase I: A wnt'er of disconnect. **Curr. Biol**. 12, R499.

74. Mao, W., Luis, E., Ross, S., Silva, J., Tan, C., Crowley, C., Chui, C., Franz, G., Senter, P., Koeppen, H., **Polakis, P**. 2004 EphB2 as a therapeutic antibody drug target for the treatment of colorectal cancer. **Cancer Res**. 64, 781-788.

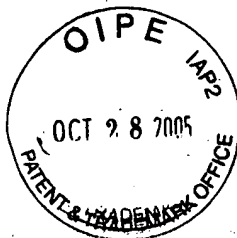
75. Shibamoto, S., Winer, J., Williams, M., **Polakis, P**. 2003 A Blockade in Wnt signaling is activated following the differentiation of F9 teratocarcinoma cells. **Exp. Cell Res**. 29211-20.

76. Zhang Y, Eberhard DA, Frantz GD, Dowd P, Wu TD, Zhou Y, Watanabe C, Luoh SM, **Polakis P**, Hillan KJ, Wood WI, Zhang Z. 2004 GEPIS--quantitative gene expression profiling in normal and cancer tissues. **Bioinformatics**, April 8



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant : Ashkenazi et al.  
App. No. : 09/903,925  
Filed : July 11, 2001  
For : SECRETED AND  
TRANSMEMBRANE  
POLYPEPTIDES AND NUCLEIC  
ACIDS ENCODING THE SAME  
Examiner : Hamud, Fozia M



Group Art Unit 1647

CERTIFICATE OF EXPRESS MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to Commissioner of Patents, Washington D.C. 20231 on:

(Date)

Commissioner of Patents  
P.O. Box 1450  
Alexandria, VA 22313-1450

**DECLARATION OF AVI ASHKENAZI, Ph.D UNDER 37 C.F.R. § 1.132**

I, Avi Ashkenazi, Ph.D. declare and say as follows: -

1. I am Director and Staff Scientist at the Molecular Oncology Department of Genentech, Inc., South San Francisco, CA 94080.
2. I joined Genentech in 1988 as a postdoctoral fellow. Since then, I have investigated a variety of cellular signal transduction mechanisms, including apoptosis, and have developed technologies to modulate such mechanisms as a means of therapeutic intervention in cancer and autoimmune disease. I am currently involved in the investigation of a series of secreted proteins over-expressed in tumors, with the aim to identify useful targets for the development of therapeutic antibodies for cancer treatment.
3. My scientific Curriculum Vitae, including my list of publications, is attached to and forms part of this Declaration (Exhibit A).
4. Gene amplification is a process in which chromosomes undergo changes to contain multiple copies of certain genes that normally exist as a single copy, and is an important factor in the pathophysiology of cancer. Amplification of certain genes (e.g., Myc or Her2/Neu)

gives cancer cells a growth or survival advantage relative to normal cells, and might also provide a mechanism of tumor cell resistance to chemotherapy or radiotherapy.

5. If gene amplification results in over-expression of the mRNA and the corresponding gene product, then it identifies that gene product as a promising target for cancer therapy, for example by the therapeutic antibody approach. Even in the absence of over-expression of the gene product, amplification of a cancer marker gene - as detected, for example, by the reverse transcriptase TaqMan<sup>®</sup> PCR or the fluorescence *in situ* hybridization (FISH) assays - is useful in the diagnosis or classification of cancer, or in predicting or monitoring the efficacy of cancer therapy. An increase in gene copy number can result not only from intrachromosomal changes but also from chromosomal aneuploidy. It is important to understand that detection of gene amplification can be used for cancer diagnosis even if the determination includes measurement of chromosomal aneuploidy. Indeed, as long as a significant difference relative to normal tissue is detected, it is irrelevant if the signal originates from an increase in the number of gene copies per chromosome and/or an abnormal number of chromosomes.

6. I understand that according to the Patent Office, absent data demonstrating that the increased copy number of a gene in certain types of cancer leads to increased expression of its product, gene amplification data are insufficient to provide substantial utility or well established utility for the gene product (the encoded polypeptide), or an antibody specifically binding the encoded polypeptide. However, even when amplification of a cancer marker gene does not result in significant over-expression of the corresponding gene product, this very absence of gene product over-expression still provides significant information for cancer diagnosis and treatment. Thus, if over-expression of the gene product does not parallel gene amplification in certain tumor types but does so in others, then parallel monitoring of gene amplification and gene product over-expression enables more accurate tumor classification and hence better determination of suitable therapy. In addition, absence of over-expression is crucial information for the practicing clinician. If a gene is amplified but the corresponding gene product is not over-expressed, the clinician accordingly will decide not to treat a patient with agents that target that gene product.

7. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information or belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so

made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful statements may jeopardize the validity of the application or any patent issued thereon.

By: Avi Ashkenazi  
Avi Ashkenazi, Ph.D.

Date: 9/15/03

## **CURRICULUM VITAE**

**Avi Ashkenazi**

July 2003

### **Personal:**

Date of birth: 29 November, 1956  
Address: 1456 Tarrytown Street, San Mateo, CA 94402  
Phone: (650) 578-9199 (home); (650) 225-1853 (office)  
Fax: (650) 225-6443 (office)  
Email: aa@gene.com

### **Education:**

1983: B.S. in Biochemistry, with honors, Hebrew University, Israel  
1986: Ph.D. in Biochemistry, Hebrew University, Israel

### **Employment:**

1983-1986: Teaching assistant, undergraduate level course in Biochemistry  
1985-1986: Teaching assistant, graduate level course on Signal Transduction  
1986 - 1988: Postdoctoral fellow, Hormone Research Dept., UCSF, and  
Developmental Biology Dept., Genentech, Inc., with J. Ramachandran  
1988 - 1989: Postdoctoral fellow, Molecular Biology Dept., Genentech, Inc.,  
with D. Capon  
1989 - 1993: Scientist, Molecular Biology Dept., Genentech, Inc.  
1994 -1996: Senior Scientist, Molecular Oncology Dept., Genentech, Inc.  
1996-1997: Senior Scientist and Interim director, Molecular Oncology Dept.,  
Genentech, Inc.  
1997-1990: Senior Scientist and preclinical project team leader, Genentech, Inc.  
1999 -2002: Staff Scientist in Molecular Oncology, Genentech, Inc.  
2002-present: Staff Scientist and Director in Molecular Oncology, Genentech, Inc.

### **Awards:**

1988: First prize, The Boehringer Ingelheim Award

## Editorial:

Editorial Board Member: Current Biology

Associate Editor, Clinical Cancer Research.

Associate Editor, Cancer Biology and Therapy.

## Refereed papers:

1. Gertler, A., Ashkenazi, A., and Madar, Z. Binding sites for human growth hormone and ovine and bovine prolactins in the mammary gland and liver of the lactating cow. *Mol. Cell. Endocrinol.* **34**, 51-57 (1984).
2. Gertler, A., Shamay, A., Cohen, N., Ashkenazi, A., Friesen, H., Levanon, A., Gorecki, M., Aviv, H., Hadari, D., and Vogel, T. Inhibition of lactogenic activities of ovine prolactin and human growth hormone (hGH) by a novel form of a modified recombinant hGH. *Endocrinology* **118**, 720-726 (1986).
3. Ashkenazi, A., Madar, Z., and Gertler, A. Partial purification and characterization of bovine mammary gland prolactin receptor. *Mol. Cell. Endocrinol.* **50**, 79-87 (1987).
4. Ashkenazi, A., Pines, M., and Gertler, A. Down-regulation of lactogenic hormone receptors in Nb2 lymphoma cells by cholera toxin. *Biochemistry Internatl.* **14**, 1065-1072 (1987).
5. Ashkenazi, A., Cohen, R., and Gertler, A. Characterization of lactogen receptors in lactogenic hormone-dependent and independent Nb2 lymphoma cell lines. *FEBS Lett.* **210**, 51-55 (1987).
6. Ashkenazi, A., Vogel, T., Barash, I., Hadari, D., Levanon, A., Gorecki, M., and Gertler, A. Comparative study on in vitro and in vivo modulation of lactogenic and somatotrophic receptors by native human growth hormone and its modified recombinant analog. *Endocrinology* **121**, 414-419 (1987).
7. Peralta, E., Winslow, J., Peterson, G., Smith, D., Ashkenazi, A., Ramachandran, J., Schimerlik, M., and Capon, D. Primary structure and biochemical properties of an M2 muscarinic receptor. *Science* **236**, 600-605 (1987).
8. Peralta, E., Ashkenazi, A., Winslow, J., Smith, D., Ramachandran, J., and Capon, D. J. Distinct primary structures, ligand-binding properties and tissue-specific expression of four human muscarinic acetylcholine receptors. *EMBO J.* **6**, 3923-3929 (1987).
9. Ashkenazi, A., Winslow, J., Peralta, E., Peterson, G., Schimerlik, M., Capon, D., and Ramachandran, J. An M2 muscarinic receptor subtype coupled to both adenylyl cyclase and phosphoinositide turnover. *Science* **238**, 672-675 (1987).

10. Pines, M., Ashkenazi, A., Cohen-Chapnik, N., Binder, L., and Gertler, A. Inhibition of the proliferation of Nb2 lymphoma cells by femtomolar concentrations of cholera toxin and partial reversal of the effect by 12-o-tetradecanoyl-phorbol-13-acetate. *J. Cell. Biochem.* **37**, 119-129 (1988).
11. Peralta, E., Ashkenazi, A., Winslow, J., Ramachandran, J., and Capon, D. Differential regulation of PI hydrolysis and adenylyl cyclase by muscarinic receptor subtypes. *Nature* **334**, 434-437 (1988).
12. Ashkenazi, A., Peralta, E., Winslow, J., Ramachandran, J., and Capon, D. Functionally distinct G proteins couple different receptors to PI hydrolysis in the same cell. *Cell* **56**, 487-493 (1989).
13. Ashkenazi, A., Ramachandran, J., and Capon, D. Acetylcholine analogue stimulates DNA synthesis in brain-derived cells via specific muscarinic acetylcholine receptor subtypes. *Nature* **340**, 146-150 (1989).
14. Lammare, D., Ashkenazi, A., Fleury, S., Smith, D., Sekaly, R., and Capon, D. The MHC-binding and gp120-binding domains of CD4 are distinct and separable. *Science* **245**, 743-745 (1989).
15. Ashkenazi, A., Presta, L., Marsters, S., Camerato, T., Rosenthal, K., Fendly, B., and Capon, D. Mapping the CD4 binding site for human immunodeficiency virus type 1 by alanine-scanning mutagenesis. *Proc. Natl. Acad. Sci. USA.* **87**, 7150-7154 (1990).
16. Chamow, S., Peers, D., Byrn, R., Mulkerrin, M., Harris, R., Wang, W., Bjorkman, P., Capon, D., and Ashkenazi, A. Enzymatic cleavage of a CD4 immunoadhesin generates crystallizable, biologically active Fd-like fragments. *Biochemistry* **29**, 9885-9891 (1990).
17. Ashkenazi, A., Smith, D., Marsters, S., Riddle, L., Gregory, T., Ho, D., and Capon, D. Resistance of primary isolates of human immunodeficiency virus type 1 to soluble CD4 is independent of CD4-gp120 binding affinity. *Proc. Natl. Acad. Sci. USA.* **88**, 7056-7060 (1991).
18. Ashkenazi, A., Marsters, S., Capon, D., Chamow, S., Figari, I., Pennica, D., Goeddel, D., Palladino, M., and Smith, D. Protection against endotoxic shock by a tumor necrosis factor receptor immunoadhesin. *Proc. Natl. Acad. Sci. USA.* **88**, 10535-10539 (1991).
19. Moore, J., McKeating, J., Huang, Y., Ashkenazi, A., and Ho, D. Virions of primary HIV-1 isolates resistant to sCD4 neutralization differ in sCD4 affinity and glycoprotein gp120 retention from sCD4-sensitive isolates. *J. Virol.* **66**, 235-243 (1992).

20. Jin, H., Oksenberg, D., Ashkenazi, A., Peroutka, S., Duncan, A., Rozmahel, R., Yang, Y., Mengod, G., Palacios, J., and O'Dowd, B. Characterization of the human 5-hydroxytryptamine<sub>1B</sub> receptor. *J. Biol. Chem.* **267**, 5735-5738 (1992).
21. Marsters, A., Frutkin, A., Simpson, N., Fendly, B. and Ashkenazi, A. Identification of cysteine-rich domains of the type 1 tumor necrosis receptor involved in ligand binding. *J. Biol. Chem.* **267**, 5747-5750 (1992).
22. Chamow, S., Kogan, T., Peers, D., Hastings, R., Byrn, R., and Ashkenazi, A. Conjugation of sCD4 without loss of biological activity via a novel carbohydrate-directed cross-linking reagent. *J. Biol. Chem.* **267**, 15916-15922 (1992).
23. Oksenberg, D., Marsters, A., O'Dowd, B., Jin, H., Havlik, S., Peroutka, S., and Ashkenazi, A. A single amino-acid difference confers major pharmacologic variation between human and rodent 5-HT<sub>1B</sub> receptors. *Nature* **360**, 161-163 (1992).
24. Haak-Frendscho, M., Marsters, S., Chamow, S., Peers, D., Simpson, N., and Ashkenazi, A. Inhibition of interferon  $\gamma$  by an interferon  $\gamma$  receptor immunoadhesin. *Immunology* **79**, 594-599 (1993).
25. Penica, D., Lam, V., Weber, R., Kohr, W., Basa, L., Spellman, M., Ashkenazi, A., Shire, S., and Goeddel, D. Biochemical characterization of the extracellular domain of the 75-kd tumor necrosis factor receptor. *Biochemistry* **32**, 3131-3138. (1993).
26. Barford, L., Zheng, Y., Kuang, W., Hart, M., Evans, T., Cerione, R., and Ashkenazi, A. Cloning and expression of a human CDC42 GTPase Activating Protein reveals a functional SH3-binding domain. *J. Biol. Chem.* **268**, 26059-26062 (1993).
27. Chamow, S., Zhang, D., Tan, X., Mhtre, S., Marsters, S., Peers, D., Byrn, R., Ashkenazi, A., and Yunghans, R. A humanized bispecific immunoadhesin-antibody that retargets CD3<sup>+</sup> effectors to kill HIV-1-infected cells. *J. Immunol.* **153**, 4268-4280 (1994).
28. Means, R., Krantz, S., Luna, J., Marsters, S., and Ashkenazi, A. Inhibition of murine erythroid colony formation in vitro by iterferon  $\gamma$  and correction by interferon  $\gamma$  receptor immunoadhesin. *Blood* **83**, 911-915 (1994).
29. Haak-Frendscho, M., Marsters, S., Mordenti, J., Gillet, N., Chen, S., and Ashkenazi, A. Inhibition of TNF by a TNF receptor immunoadhesin: comparison with an anti-TNF mAb. *J. Immunol.* **152**, 1347-1353 (1994).

30. Chamow, S., Kogan, T., Venuti, M., Gadek, T., Peers, D., Mordenti, J., Shak, S., and Ashkenazi, A. Modification of CD4 immunoadhesin with monomethoxy-PEG aldehyde via reductive alkylation. *Bioconj. Chem.* **5**, 133-140 (1994).
31. Jin, H., Yang, R., Marsters, S., Bunting, S., Wurm, F., Chamow, S., and Ashkenazi, A. Protection against rat endotoxic shock by p55 tumor necrosis factor (TNF) receptor immunoadhesin: comparison to anti-TNF monoclonal antibody. *J. Infect. Diseases* **170**, 1323-1326 (1994).
32. Beck, J., Marsters, S., Harris, R., Ashkenazi, A., and Chamow, S. Generation of soluble interleukin-1 receptor from an immunoadhesin by specific cleavage. *Mol. Immunol.* **31**, 1335-1344 (1994).
33. Pitti, B., Marsters, M., Haak-Frendscho, M., Osaka, G., Mordenti, J., Chamow, S., and Ashkenazi, A. Molecular and biological properties of an interleukin-1 receptor immunoadhesin. *Mol. Immunol.* **31**, 1345-1351 (1994).
34. Oksenberg, D., Havlik, S., Peroutka, S., and Ashkenazi, A. The third intracellular loop of the 5-HT<sub>2</sub> receptor specifies effector coupling. *J. Neurochem.* **64**, 1440-1447 (1995).
35. Bach, E., Szabo, S., Dighe, A., Ashkenazi, A., Aguet, M., Murphy, K., and Schreiber, R. Ligand-induced autoregulation of IFN- $\gamma$  receptor  $\beta$  chain expression in T helper cell subsets. *Science* **270**, 1215-1218 (1995).
36. Jin, H., Yang, R., Marsters, S., Ashkenazi, A., Bunting, S., Marra, M., Scott, R., and Baker, J. Protection against endotoxic shock by bactericidal/permeability-increasing protein in rats. *J. Clin. Invest.* **95**, 1947-1952 (1995).
37. Marsters, S., Penica, D., Bach, E., Schreiber, R., and Ashkenazi, A. Interferon  $\gamma$  signals via a high-affinity multisubunit receptor complex that contains two types of polypeptide chain. *Proc. Natl. Acad. Sci. USA.* **92**, 5401-5405 (1995).
38. Van Zee, K., Moldawer, L., Oldenburg, H., Thompson, W., Stackpole, S., Montegut, W., Rogy, M., Meschter, C., Gallati, H., Schiller, C., Richter, W., Loetcher, H., Ashkenazi, A., Chamow, S., Wurm, F., Calvano, S., Lowry, S., and Lesslauer, W. Protection against lethal *E. coli* bacteremia in baboons by pretreatment with a 55-kDa TNF receptor-Ig fusion protein, Ro45-2081. *J. Immunol.* **156**, 2221-2230 (1996).
39. Pitti, R., Marsters, S., Ruppert, S., Donahue, C., Moore, A., and Ashkenazi, A. Induction of apoptosis by Apo-2 Ligand, a new member of the tumor necrosis factor cytokine family. *J. Biol. Chem.* **271**, 12687-12690 (1996).



40. Marsters, S., Pitti, R., Donahue, C., Rupert, S., Bauer, K., and Ashkenazi, A. Activation of apoptosis by Apo-2 ligand is independent of FADD but blocked by CrmA. *Curr. Biol.* 6, 1669-1676 (1996).
41. Marsters, S., Skubatch, M., Gray, C., and Ashkenazi, A. Herpesvirus entry mediator, a novel member of the tumor necrosis factor receptor family, activates the NF- $\kappa$ B and AP-1 transcription factors. *J. Biol. Chem.* 272, 14029-14032 (1997).
42. Sheridan, J., Marsters, S., Pitti, R., Gurney, A., Skubatch, M., Baldwin, D., Ramakrishnan, L., Gray, C., Baker, K., Wood, W.I., Goddard, A., Godowski, P., and Ashkenazi, A. Control of TRAIL-induced apoptosis by a family of signaling and decoy receptors. *Science* 277, 818-821 (1997).
43. Marsters, S., Sheridan, J., Pitti, R., Gurney, A., Skubatch, M., Baldwin, D., Huang, A., Yuan, J., Goddard, A., Godowski, P., and Ashkenazi, A. A novel receptor for Apo2L/TRAIL contains a truncated death domain. *Curr. Biol.* 7, 1003-1006 (1997).
44. Marsters, A., Sheridan, J., Pitti, R., Brush, J., Goddard, A., and Ashkenazi, A. Identification of a ligand for the death-domain-containing receptor Apo3. *Curr. Biol.* 8, 525-528 (1998).
45. Rieger, J., Naumann, U., Glaser, T., Ashkenazi, A., and Weller, M. Apo2 ligand: a novel weapon against malignant glioma? *FEBS Lett.* 427, 124-128 (1998).
46. Pender, S., Fell, J., Chamow, S., Ashkenazi, A., and MacDonald, T. A p55 TNF receptor immunoadhesin prevents T cell mediated intestinal injury by inhibiting matrix metalloproteinase production. *J. Immunol.* 160, 4098-4103 (1998).
47. Pitti, R., Marsters, S., Lawrence, D., Roy, Kischkel, F., M., Dowd, P., Huang, A., Donahue, C., Sherwood, S., Baldwin, D., Godowski, P., Wood, W., Gurney, A., Hillan, K., Cohen, R., Goddard, A., Botstein, D., and Ashkenazi, A. Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer. *Nature* 396, 699-703 (1998).
48. Mori, S., Marakami-Mori, K., Nakamura, S., Ashkenazi, A., and Bonavida, B. Sensitization of AIDS Kaposi's sarcoma cells to Apo-2 ligand-induced apoptosis by actinomycin D. *J. Immunol.* 162, 5616-5623 (1999).
49. Gurney, A. Marsters, S., Huang, A., Pitti, R., Mark, M., Baldwin, D., Gray, A., Dowd, P., Brush, J., Heldens, S., Schow, P., Goddard, A., Wood, W., Baker, K., Godowski, P., and Ashkenazi, A. Identification of a new member of the tumor necrosis factor family and its receptor, a human ortholog of mouse GITR. *Curr. Biol.* 9, 215-218 (1999).

50. Ashkenazi, A., Pai, R., Fong, s., Leung, S., Lawrence, D., Marsters, S., Blackie, C., Chang, L., McMurtrey, A., Hebert, A., DeForge, L., Khoumenis, I., Lewis, D., Harris, L., Bussiere, J., Koeppen, H., Shahrokh, Z., and Schwall, R. Safety and anti-tumor activity of recombinant soluble Apo2 ligand. *J. Clin. Invest.* **104**, 155-162 (1999).
51. Chuntharapai, A., Gibbs, V., Lu, J., Ow, A., Marsters, S., Ashkenazi, A., De Vos, A., Kim, K.J. Determination of residues involved in ligand binding and signal transmissiion in the human IFN- $\alpha$  receptor 2. *J. Immunol.* **163**, 766-773 (1999).
52. Johnsen, A.-C., Haux, J., Steinkjer, B., Nonstad, U., Egeberg, K., Sundan, A., Ashkenazi, A., and Espevik, T. Regulation of Apo2L/TRAIL expression in NK cells – involvement in NK cell-mediated cytotoxicity. *Cytokine* **11**, 664-672 (1999).
53. Roth, W., Isenmann, S., Naumann, U., Kugler, S., Bahr, M., Dichgans, J., Ashkenazi, A., and Weller, M. Eradication of intracranial human malignant glioma xenografts by Apo2L/TRAIL. *Biochem. Biophys. Res. Commun.* **265**, 479-483 (1999).
54. Hymowitz, S.G., Christinger, H.W., Fuh, G., Ultsch, M., O'Connell, M., Kelley, R.F., Ashkenazi, A. and de Vos, A.M. Triggering Cell Death: The Crystal Structure of Apo2L/TRAIL in a Complex with Death Receptor 5. *Molec. Cell* **4**, 563–571 (1999).
55. Hymowitz, S.G., O'Connel, M.P., Utsch, M.H., Hurst, A., Totpal, K., Ashkenazi, A., de Vos, A.M., Kelley, R.F. A unique zinc-binding site revealed by a high-resolution X-ray structure of homotrimeric Apo2L/TRAIL. *Biochemistry* **39**, 633-640 (2000).
56. Zhou, Q., Fukushima, P., DeGraff, W., Mitchell, J.B., Stetler-Stevenson, M., Ashkenazi, A., and Steeg, P.S. Radiation and the Apo2L/TRAIL apoptotic pathway preferentially inhibit the colonization of premalignant human breast cancer cells overexpressing cyclin D1. *Cancer Res.* **60**, 2611-2615 (2000).
57. Kischkel, F.C., Lawrence, D. A., Chuntharapai, A., Schow, P., Kim, J., and Ashkenazi, A. Apo2L/TRAIL-dependent recruitment of endogenous FADD and Caspase-8 to death receptors 4 and 5. *Immunity* **12**, 611-620 (2000).
58. Yan, M., Marsters, S.A., Grewal, I.S., Wang, H., \*Ashkenazi, A., and \*Dixit, V.M. Identification of a receptor for BlyS demonstrates a crucial role in humoral immunity. *Nature Immunol.* **1**, 37-41 (2000).

59. Marsters, S.A., Yan, M., Pitti, R.M., Haas, P.E., Dixit, V.M., and Ashkenazi, A. Interaction of the TNF homologues BLyS and APRIL with the TNF receptor homologues BCMA and TACI. *Curr. Biol.* **10**, 785-788 (2000).
60. Kischkel, F.C., and Ashkenazi, A. Combining enhanced metabolic labeling with immunoblotting to detect interactions of endogenous cellular proteins. *Biotechniques* **29**, 506-512 (2000).
61. Lawrence, D., Shahrokh, Z., Marsters, S., Achilles, K., Shih, D. Mounho, B., Hillan, K., Totpal, K. DeForge, L., Schow, P., Hooley, J., Sherwood, S., Pai, R., Leung, S., Khan, L., Gliniak, B., Bussiere, J., Smith, C., Strom, S., Kelley, S., Fox, J., Thomas, D., and Ashkenazi, A. Differential hepatocyte toxicity of recombinant Apo2L/TRAIL versions. *Nature Med.* **7**, 383-385 (2001).
62. Chuntharapai, A., Dodge, K., Grimmer, K., Schroeder, K., Marsters, S.A., Koeppen, H., Ashkenazi, A., and Kim, K.J. Isotype-dependent inhibition of tumor growth in vivo by monoclonal antibodies to death receptor 4. *J. Immunol.* **166**, 4891-4898 (2001).
63. Pollack, I.F., Erff, M., and Ashkenazi, A. Direct stimulation of apoptotic signaling by soluble Apo2L/tumor necrosis factor-related apoptosis-inducing ligand leads to selective killing of glioma cells. *Clin. Cancer Res.* **7**, 1362-1369 (2001).
64. Wang, H., Marsters, S.A., Baker, T., Chan, B., Lee, W.P., Fu, L., Tumas, D., Yan, M., Dixit, V.M., \*Ashkenazi, A., and \*Grewal, I.S. TACI-ligand interactions are required for T cell activation and collagen-induced arthritis in mice. *Nature Immunol.* **2**, 632-637 (2001).
65. Kischkel, F.C., Lawrence, D. A., Tinel, A., Virmani, A., Schow, P., Gazdar, A., Blenis, J., Arnott, D., and Ashkenazi, A. Death receptor recruitment of endogenous caspase-10 and apoptosis initiation in the absence of caspase-8. *J. Biol. Chem.* **276**, 46639-46646 (2001).
66. LeBlanc, H., Lawrence, D.A., Varfolomeev, E., Totpal, K., Morlan, J., Schow, P., Fong, S., Schwall, R., Sinicropi, D., and Ashkenazi, A. Tumor cell resistance to death receptor induced apoptosis through mutational inactivation of the proapoptotic Bcl-2 homolog Bax. *Nature Med.* **8**, 274-281 (2002).
67. Miller, K., Meng, G., Liu, J., Hurst, A., Hsei, V., Wong, W-L., Ekert, R., Lawrence, D., Sherwood, S., DeForge, L., Gaudreault, G., Keller, G., Sliwkowski, M., Ashkenazi, A., and Presta, L. Design, Construction, and analyses of multivalent antibodies. *J. Immunol.* **170**, 4854-4861 (2003).

68. Varfolomeev, E., Kischkel, F., Martin, F., Wanh, H., Lawrence, D., Olsson, C., Tom, L., Erickson, S., French, D., Schow, P., Grewal, I. and Ashkenazi, A. Immune system development in APRIL knockout mice. Submitted.

**Review articles:**

1. Ashkenazi, A., Peralta, E., Winslow, J., Ramachandran, J., and Capon, D., J. Functional role of muscarinic acetylcholine receptor subtype diversity. *Cold Spring Harbor Symposium on Quantitative Biology*. **LIII**, 263-272 (1988).
2. Ashkenazi, A., Peralta, E., Winslow, J., Ramachandran, J., and Capon, D. Functional diversity of muscarinic receptor subtypes in cellular signal transduction and growth. *Trends Pharmacol. Sci.* Dec Supplement, 12-21 (1989).
3. Chamow, S., Duliege, A., Ammann, A., Kahn, J., Allen, D., Eichberg, J., Byrn, R., Capon, D., Ward, R., and Ashkenazi, A. CD4 immunoadhesins in anti-HIV therapy: new developments. *Int. J. Cancer* Supplement 7, 69-72 (1992).
4. Ashkenazi, A., Capon, and D. Ward, R. Immunoadhesins. *Int. Rev. Immunol.* **10**, 217-225 (1993).
5. Ashkenazi, A., and Peralta, E. Muscarinic Receptors. In *Handbook of Receptors and Channels*. (S. Peroutka, ed.), CRC Press, Boca Raton, Vol. I, p. 1-27, (1994).
6. Krantz, S. B., Means, R. T., Jr., Lina, J., Marsters, S. A., and Ashkenazi, A. Inhibition of erythroid colony formation in vitro by gamma interferon. In *Molecular Biology of Hematopoiesis* (N. Abraham, R. Shadduck, A. Levine F. Takaku, eds.) Intercept Ltd. Paris, Vol. 3, p. 135-147 (1994).
7. Ashkenazi, A. Cytokine neutralization as a potential therapeutic approach for SIRS and shock. *J. Biotechnology in Healthcare* **1**, 197-206 (1994).
8. Ashkenazi, A., and Chamow, S. M. Immunoadhesins: an alternative to human monoclonal antibodies. *Immunomethods: A companion to Methods in Enzimology* **8**, 104-115 (1995).
9. Chamow, S., and Ashkenazi, A. Immunoadhesins: Principles and Applications. *Trends Biotech.* **14**, 52-60 (1996).
10. Ashkenazi, A., and Chamow, S. M. Immunoadhesins as research tools and therapeutic agents. *Curr. Opin. Immunol.* **9**, 195-200 (1997).
11. Ashkenazi, A., and Dixit, V. Death receptors: signaling and modulation. *Science* **281**, 1305-1308 (1998).
12. Ashkenazi, A., and Dixit, V. Apoptosis control by death and decoy receptors. *Curr. Opin. Cell. Biol.* **11**, 255-260 (1999).

13. Ashkenazi, A. Chapters on Apo2L/TRAIL; DR4, DR5, DcR1, DcR2; and DcR3. Online Cytokine Handbook ([www.apnet.com/cytokinereference/](http://www.apnet.com/cytokinereference/)).
14. Ashkenazi, A. Targeting death and decoy receptors of the tumor necrosis factor superfamily. *Nature Rev. Cancer* 2, 420-430 (2002).
15. LeBlanc, H. and Ashkenazi, A. Apoptosis signaling by Apo2L/TRAIL. *Cell Death and Differentiation* 10, 66-75 (2003).
16. Almasan, A. and Ashkenazi, A. Apo2L/TRAIL: apoptosis signaling, biology, and potential for cancer therapy. *Cytokine and Growth Factor Reviews* 14, 337-348 (2003).

**Book:**

Antibody Fusion Proteins (Chamow, S., and Ashkenazi, A., eds., John Wiley and Sons Inc.) (1999).

**Talks:**

1. Resistance of primary HIV isolates to CD4 is independent of CD4-gp120 binding affinity. UCSD Symposium, HIV Disease: Pathogenesis and Therapy. Greenelefe, FL, March 1991.
2. Use of immuno-hybrids to extend the half-life of receptors. IBC conference on Biopharmaceutical Half-life Extension. New Orleans, LA, June 1992.
3. Results with TNF receptor Immunoadhesins for the Treatment of Sepsis. IBC conference on Endotoxemia and Sepsis. Philadelphia, PA, June 1992.
4. Immunoadhesins: an alternative to human antibodies. IBC conference on Antibody Engineering. San Diego, CA, December 1993.
5. Tumor necrosis factor receptor: a potential therapeutic for human septic shock. American Society for Microbiology Meeting, Atlanta, GA, May 1993.
6. Protective efficacy of TNF receptor immunoadhesin vs anti-TNF monoclonal antibody in a rat model for endotoxic shock. 5th International Congress on TNF. Asilomar, CA, May 1994.
7. Interferon- $\gamma$  signals via a multisubunit receptor complex that contains two types of polypeptide chain. American Association of Immunologists Conference. San Francisco, CA, July 1995.
8. Immunoadhesins: Principles and Applications. Gordon Research Conference on Drug Delivery in Biology and Medicine. Ventura, CA, February 1996.

9. Apo-2 Ligand, a new member of the TNF family that induces apoptosis in tumor cells. Cambridge Symposium on TNF and Related Cytokines in Treatment of Cancer. Hilton-Head, NC, March 1996.
10. Induction of apoptosis by Apo2 Ligand. American Society for Biochemistry and Molecular Biology, Symposium on Growth Factors and Cytokine Receptors. New Orleans, LA, June, 1996.
11. Apo2 ligand, an extracellular trigger of apoptosis. 2nd Clontech Symposium, Palo Alto, CA, October 1996.
12. Regulation of apoptosis by members of the TNF ligand and receptor families. Stanford University School of Medicine, Palo Alto, CA, December 1996.
13. Apo-3: a novel receptor that regulates cell death and inflammation. 4th International Congress on Immune Consequences of Trauma, Shock, and Sepsis. Munich, Germany, March 1997.
14. New members of the TNF ligand and receptor families that regulate apoptosis, inflammation, and immunity. UCLA School of Medicine, LA, CA, March 1997.
15. Immunoadhesins: an alternative to monoclonal antibodies. 5th World Conference on Bispecific Antibodies. Volendam, Holland, June 1997.
16. Control of Apo2L signaling. Cold Spring Harbor Laboratory Symposium on Programmed Cell Death. Cold Spring Harbor, New York. September, 1997.
17. Chairman and speaker, Apoptosis Signaling session. IBC's 4th Annual Conference on Apoptosis. San Diego, CA., October 1997.
18. Control of Apo2L signaling by death and decoy receptors. American Association for the Advancement of Science. Philadelphia, PA, February 1998.
19. Apo2 ligand and its receptors. American Society of Immunologists. San Francisco, CA, April 1998.
20. Death receptors and ligands. 7th International TNF Congress. Cape Cod, MA, May 1998.
21. Apo2L as a potential therapeutic for cancer. UCLA School of Medicine. LA, CA, June 1998.
22. Apo2L as a potential therapeutic for cancer. Gordon Research Conference on Cancer Chemotherapy. New London, NH, July 1998.
23. Control of apoptosis by Apo2L. Endocrine Society Conference, Stevenson, WA, August 1998.
24. Control of apoptosis by Apo2L. International Cytokine Society Conference, Jerusalem, Israel, October 1998.

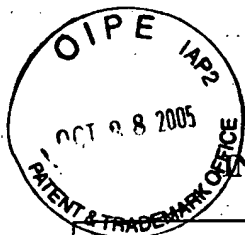
25. Apoptosis control by death and decoy receptors. American Association for Cancer Research Conference, Whistler, BC, Canada, March 1999.
26. Apoptosis control by death and decoy receptors. American Society for Biochemistry and Molecular Biology Conference, San Francisco, CA, May 1999.
27. Apoptosis control by death and decoy receptors. Gordon Research Conference on Apoptosis, New London, NH, June 1999.
28. Apoptosis control by death and decoy receptors. Arthritis Foundation Research Conference, Alexandria GA, Aug 1999.
29. Safety and anti-tumor activity of recombinant soluble Apo2L/TRAIL. Cold Spring Harbor Laboratory Symposium on Programmed Cell Death. . Cold Spring Harbor, NY, September 1999.
30. The Apo2L/TRAIL system: therapeutic potential. American Association for Cancer Research, Lake Tahoe, NV, Feb 2000.
31. Apoptosis and cancer therapy. Stanford University School of Medicine, Stanford, CA, Mar 2000.
32. Apoptosis and cancer therapy. University of Pennsylvania School of Medicine, Philadelphia, PA, Apr 2000.
33. Apoptosis signaling by Apo2L/TRAIL. International Congress on TNF. Trondheim, Norway, May 2000.
34. The Apo2L/TRAIL system: therapeutic potential. Cap-CURE summit meeting. Santa Monica, CA, June 2000.
35. The Apo2L/TRAIL system: therapeutic potential. MD Anderson Cancer Center. Houston, TX, June 2000.
36. Apoptosis signaling by Apo2L/TRAIL. The Protein Society, 14<sup>th</sup> Symposium. San Diego, CA, August 2000.
37. Anti-tumor activity of Apo2L/TRAIL. AAPS annual meeting. Indianapolis, IN Aug 2000.
38. Apoptosis signaling and anti-cancer potential of Apo2L/TRAIL. Cancer Research Institute, UC San Francisco, CA, September 2000.
39. Apoptosis signaling by Apo2L/TRAIL. Kenote address, TNF family Minisymposium, NIH. Bethesda, MD, September 2000.
40. Death receptors: signaling and modulation. Keystone symposium on the Molecular basis of cancer. Taos, NM, Jan 2001.
41. Preclinical studies of Apo2L/TRAIL in cancer. Symposium on Targeted therapies in the treatment of lung cancer. Aspen, CO, Jan 2001.

42. Apoptosis signaling by Apo2L/TRAIL. Weizmann Institute of Science, Rehovot, Israel, March 2001.
43. Apo2L/TRAIL: Apoptosis signaling and potential for cancer therapy. Weizmann Institute of Science, Rehovot, Israel, March 2001.
44. Targeting death receptors in cancer with Apo2L/TRAIL. Cell Death and Disease conference, North Falmouth, MA, Jun 2001.
45. Targeting death receptors in cancer with Apo2L/TRAIL. Biotechnology Organization conference, San Diego, CA, Jun 2001.
46. Apo2L/TRAIL signaling and apoptosis resistance mechanisms. Gordon Research Conference on Apoptosis, Oxford, UK, July 2001.
47. Apo2L/TRAIL signaling and apoptosis resistance mechanisms. Cleveland Clinic Foundation, Cleveland, OH, Oct 2001.
48. Apoptosis signaling by death receptors: overview. International Society for Interferon and Cytokine Research conference, Cleveland, OH, Oct 2001.
49. Apoptosis signaling by death receptors. American Society of Nephrology Conference. San Francisco, CA, Oct 2001.
50. Targeting death receptors in cancer. Apoptosis: commercial opportunities. San Diego, CA, Apr 2002.
51. Apo2L/TRAIL signaling and apoptosis resistance mechanisms. Kimmel Cancer Research Center, Johns Hopkins University, Baltimore MD. May 2002.
52. Apoptosis control by Apo2L/TRAIL. (Keynote Address) University of Alabama Cancer Center Retreat, Birmingham, Ab. October 2002.
53. Apoptosis signaling by Apo2L/TRAIL. (Session co-chair) TNF international conference. San Diego, CA. October 2002.
54. Apoptosis signaling by Apo2L/TRAIL. Swiss Institute for Cancer Research (ISREC). Lausanne, Switzerland. Jan 2003.
55. Apoptosis induction with Apo2L/TRAIL. Conference on New Targets and Innovative Strategies in Cancer Treatment. Monte Carlo. February 2003.
56. Apoptosis signaling by Apo2L/TRAIL. Hermelin Brain Tumor Center Symposium on Apoptosis. Detroit, MI. April 2003.
57. Targeting apoptosis through death receptors. Sixth Annual Conference on Targeted Therapies in the Treatment of Breast Cancer. Kona, Hawaii. July 2003.
58. Targeting apoptosis through death receptors. Second International Conference on Targeted Cancer Therapy. Washington, DC. Aug 2003.

**Issued Patents:**



1. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 5,329,028 (Jul 12, 1994).
2. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 5,605,791 (Feb 25, 1997).
3. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 5,889,155 (Jul 27, 1999).
4. Ashkenazi, A., APO-2 Ligand. US patent 6,030,945 (Feb 29, 2000).
5. Ashkenazi, A., Chuntharapai, A., Kim, J., APO-2 ligand antibodies. US patent 6,046,048 (Apr 4, 2000).
6. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 6,124,435 (Sep 26, 2000).
7. Ashkenazi, A., Chuntharapai, A., Kim, J., Method for making monoclonal and cross-reactive antibodies. US patent 6,252,050 (Jun 26, 2001).
8. Ashkenazi, A. APO-2 Receptor. US patent 6,342,369 (Jan 29, 2002).
9. Ashkenazi, A. Fong, S., Goddard, A., Gurney, A., Napier, M., Tumas, D., Wood, W. A-33 polypeptides. US patent 6,410,708 (Jun 25, 2002).
10. Ashkenazi, A. APO-3 Receptor. US patent 6,462,176 B1 (Oct 8, 2002).
11. Ashkenazi, A. APO-2LI and APO-3 polypeptide antibodies. US patent 6,469,144 B1 (Oct 22, 2002).
12. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 6,582,928B1 (Jun 24, 2003).



PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: Ashkenazi et al.	Group Art Unit: 1647
Serial No.: 09/903,925	Examiner: Fozia Hamid
Filed: July 11, 2001	<b>CERTIFICATE OF MAILING</b> I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Assistant Commissioner of Patents, Washington, D.C. 20231 on
For: SECRETED AND TRANSMEMBRANE POLYPEPTIDES AND NUCLEIC ACIDS	Date

**DECLARATION OF AUDREY D. GODDARD, Ph.D UNDER 37 C.F.R. § 1.132**

Assistant Commissioner of Patents  
Washington, D.C. 20231

Sir:

I, Audrey D. Goddard, Ph.D. do hereby declare and say as follows:

1. I am a Senior Clinical Scientist at the Experimental Medicine/BioOncology, Medical Affairs Department of Genentech, Inc., South San Francisco, California 94080.
2. Between 1993 and 2001, I headed the DNA Sequencing Laboratory at the Molecular Biology Department of Genentech, Inc. During this time, my responsibilities included the identification and characterization of genes contributing to the oncogenic process, and determination of the chromosomal localization of novel genes.
3. My scientific Curriculum Vitae, including my list of publications, is attached to and forms part of this Declaration (Exhibit A).

Serial No.: \*

Filed: \*

4. I am familiar with a variety of techniques known in the art for detecting and quantifying the amplification of oncogenes in cancer, including the quantitative TaqMan PCR (i.e., "gene amplification") assay described in the above captioned patent application.

5. The TaqMan PCR assay is described, for example, in the following scientific publications: Higuchi *et al.*, Biotechnology 10:413-417 (1992) (Exhibit B); Livak *et al.*, PCR Methods Appl., 4:357-362 (1995) (Exhibit C) and Heid *et al.*, Genome Res. 6:986-994 (1996) (Exhibit D). Briefly, the assay is based on the principle that successful PCR yields a fluorescent signal due to Taq DNA polymerase-mediated exonuclease digestion of a fluorescently labeled oligonucleotide that is homologous to a sequence between two PCR primers. The extent of digestion depends directly on the amount of PCR, and can be quantified accurately by measuring the increment in fluorescence that results from decreased energy transfer. This is an extremely sensitive technique, which allows detection in the exponential phase of the PCR reaction and, as a result, leads to accurate determination of gene copy number.

6. The quantitative fluorescent TaqMan PCR assay has been extensively and successfully used to characterize genes involved in cancer development and progression. Amplification of protooncogenes has been studied in a variety of human tumors, and is widely considered as having etiological, diagnostic and prognostic significance. This use of the quantitative TaqMan PCR assay is exemplified by the following scientific publications: Pennica *et al.*, Proc. Natl. Acad. Sci. USA 95(25):14717-14722 (1998) (Exhibit E); Pitti *et al.*, Nature 396(6712):699-703 (1998) (Exhibit F) and Bieche *et al.*, Int. J. Cancer 78:661-666 (1998) (Exhibit G), the first two of which I am co-author. In particular, Pennica *et al.* have used the quantitative TaqMan PCR assay to study relative gene amplification of WISP and c-myc in various cell lines, colorectal tumors and normal mucosa. Pitti *et al.* studied the genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer, using the quantitative TaqMan PCR assay. Bieche *et al.* used the assay to study gene amplification in breast cancer.

Serial No.: \*

Filed: \*

7. It is my personal experience that the quantitative TaqMan PCR technique is technically sensitive enough to detect at least a 2-fold increase in gene copy number relative to control. It is further my considered scientific opinion that an at least 2-fold increase in gene copy number in a tumor tissue sample relative to a normal (i.e., non-tumor) sample is significant and useful in that the detected increase in gene copy number in the tumor sample relative to the normal sample serves as a basis for using relative gene copy number as quantitated by the TaqMan PCR technique as a diagnostic marker for the presence or absence of tumor in a tissue sample of unknown pathology. Accordingly, a gene identified as being amplified at least 2-fold by the quantitative TaqMan PCR assay in a tumor sample relative to a normal sample is useful as a marker for the diagnosis of cancer, for monitoring cancer development and/or for measuring the efficacy of cancer therapy.

8. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true. I declare that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Jan. 16, 2003

Date

Audrey D. Goddard

Audrey D. Goddard, Ph.D.

**AUDREY D. GODDARD, Ph.D.**

Genentech, Inc.  
1 DNA Way  
South San Francisco, CA, 94080  
650.225.6429  
goddarda@gene.com

110 Congo St.  
San Francisco, CA, 94131  
415.841.9154  
415.819.2247 (mobile)  
agoddard@pacbell.net

**PROFESSIONAL EXPERIENCE**

**Genentech, Inc.**  
**South San Francisco, CA**

**1993-present**

**2001 - present      Senior Clinical Scientist**  
Experimental Medicine / BioOncology, Medical Affairs

**Responsibilities:**

- *Companion diagnostic oncology products*
- *Acquisition of clinical samples from Genentech's clinical trials for translational research*
- *Translational research using clinical specimen and data for drug development and diagnostics*
- *Member of Development Science Review Committee, Diagnostic Oversight Team, 21 CFR Part 11 Subteam*

**Interests:**

- *Ethical and legal implications of experiments with clinical specimens and data*
- *Application of pharmacogenomics in clinical trials*

**1998 - 2001      Senior Scientist**

Head of the DNA Sequencing Laboratory, Molecular Biology Department, Research

**Responsibilities:**

- *Management of a laboratory of up to nineteen –including postdoctoral fellow, associate scientist, senior research associate and research assistants/associate levels*
- *Management of a \$750K budget*
- *DNA sequencing core facility supporting a 350+ person research facility.*
- *DNA sequencing for high throughput gene discovery, - ESTs, cDNAs, and constructs*
- *Genomic sequence analysis and gene identification*
- *DNA sequence and primary protein analysis*

**Research:**

- *Chromosomal localization of novel genes*
- *Identification and characterization of genes contributing to the oncogenic process*
- *Identification and characterization of genes contributing to inflammatory diseases*
- *Design and development of schemes for high throughput genomic DNA sequence analysis*
- *Candidate gene prediction and evaluation*

**1993 - 1998      Scientist**

Head of the DNA Sequencing Laboratory, Molecular Biology Department, Research

**Responsibilities**

- *DNA sequencing core facility supporting a 350+ person research facility*
- *Assumed responsibility for a pre-existing team of five technicians and expanded the group into fifteen, introducing a level of middle management and additional areas of research*
- *Participated in the development of the basic plan for high throughput secreted protein discovery program – sequencing strategies, data analysis and tracking, database design*
- *High throughput EST and cDNA sequencing for new gene identification.*
- *Design and implementation of analysis tools required for high throughput gene identification.*
- *Chromosomal localization of genes encoding novel secreted proteins.*

**Research:**

- *Genomic sequence scanning for new gene discovery.*
- *Development of signal peptide selection methods.*
- *Evaluation of candidate disease genes.*
- *Growth hormone receptor gene SNPs in children with Idiopathic short stature*

**Imperial Cancer Research Fund  
London, UK with Dr. Ellen Solomon**

**1989-1992**

**6/89 – 12/92 Postdoctoral Fellow**

- *Cloning and characterization of the genes fused at the acute promyelocytic leukemia translocation breakpoints on chromosomes 17 and 15.*
- *Prepared a successfully funded European Union multi-center grant application*

**McMaster University  
Hamilton, Ontario, Canada with Dr. G. D. Sweeney**

**1983**

**5/83 – 8/83: NSERC Summer Student**

- *In vitro* metabolism of  $\beta$ -naphthoflavone in C57Bl/6J and DBA mice

**EDUCATION**

**Ph.D.**

"Phenotypic and genotypic effects of mutations in the human retinoblastoma gene."

**Supervisor:** Dr. R. A. Phillips

University of Toronto  
Toronto, Ontario, Canada.  
Department of Medical  
Biophysics.

1989

**Honours B.Sc**

"The *in vitro* metabolism of the cytochrome P-448 inducer  $\beta$ -naphthoflavone in C57BL/6J mice."

**Supervisor:** Dr. G. D. Sweeney

McMaster University,  
Hamilton, Ontario, Canada.  
Department of Biochemistry

1983

## ACADEMIC AWARDS

Imperial Cancer Research Fund Postdoctoral Fellowship	1989-1992
Medical Research Council Studentship	1983-1988
NSERC Undergraduate Summer Research Award	1983
Society of Chemical Industry Merit Award (Hons. Biochem.)	1983
Dr. Harry Lyman Hooker Scholarship	1981-1983
J.L.W. Gill Scholarship	1981-1982
Business and Professional Women's Club Scholarship	1980-1981
Wyerhauser Foundation Scholarship	1979-1980

## INVITED PRESENTATIONS

Genentech's gene discovery pipeline: High throughput identification, cloning and characterization of novel genes. Functional Genomics: From Genome to Function, Litchfield Park, AZ, USA. October 2000

High throughput identification, cloning and characterization of novel genes. G2K:Back to Science, Advances in Genome Biology and Technology I. Marco Island, FL, USA. February 2000

Quality control in DNA Sequencing: The use of Phred and Phrap. Bay Area Sequencing Users Meeting, Berkeley, CA, USA. April 1999

High throughput secreted protein identification and cloning. Tenth International Genome Sequencing and Analysis Conference, Miami, FL, USA. September 1998

The evolution of DNA sequencing: The Genentech perspective. Bay Area Sequencing Users Meeting, Berkeley, CA, USA. May 1998

Partial Growth Hormone Insensitivity: The role of GH-receptor mutations in Idiopathic Short Stature. Tenth Annual National Cooperative Growth Study Investigators Meeting, San Francisco, CA, USA. October, 1996

Growth hormone (GH) receptor defects are present in selected children with non-GH-deficient short stature: A molecular basis for partial GH-insensitivity. 76<sup>th</sup> Annual Meeting of The Endocrine Society, Anaheim, CA, USA. June 1994

A previously uncharacterized gene, myl, is fused to the retinoic acid receptor alpha gene in acute promyelocytic leukemia. XV International Association for Comparative Research on Leukemia and Related Disease, Padua, Italy. October 1991

## PATENTS

Goddard A, Godowski PJ, Gurney AL. NL2 Tie ligand homologue polypeptide. Patent Number: 6,455,496. Date of Patent: Sept. 24, 2002.

**Goddard A**, Godowski PJ and Gurney AL. NL3 Tie ligand homologue nucleic acids. Patent Number: 6,426,218. Date of Patent: July 30, 2002.

Godowski P, Gurney A, Hillan KJ, Botstein D, **Goddard A**, Roy M, Ferrara N, Tumas D, Schwall R. NL4 Tie ligand homologue nucleic acid. Patent Number: 6,4137,770. Date of Patent: July 2, 2002.

Ashkenazi A, Fong S, **Goddard A**, Gurney AL, Napier MA, Tumas D, Wood WI. Nucleic acid encoding A-33 related antigen poly peptides. Patent Number: 6,410,708. Date of Patent: Jun. 25, 2002.

Botstein DA, Cohen RL, **Goddard AD**, Gurney AL, Hillan KJ, Lawrence DA, Levine AJ, Pennica D, Roy MA and Wood WI. WISP polypeptides and nucleic acids encoding same. Patent Number: 6,387,657. Date of Patent: May 14, 2002.

**Goddard A**, Godowski PJ and Gurney AL. Tie ligands. Patent Number: 6,372,491. Date of Patent: April 16, 2002.

Godowski PJ, Gurney AL, **Goddard A** and Hillan K. TIE ligand homologue antibody. Patent Number: 6,350,450. Date of Patent: Feb. 26, 2002.

Fong S, Ferrara N, **Goddard A**, Godowski PJ, Gurney AL, Hillan K and Williams PM. Tie receptor tyrosine kinase ligand homologues. Patent Number: 6,348,351. Date of Patent: Feb. 19, 2002.

**Goddard A**, Godowski PJ and Gurney AL. Ligand homologues. Patent Number: 6,348,350. Date of Patent: Feb. 19, 2002.

Attie KM, Carlsson LMS, Gesundheit N and **Goddard A**. Treatment of partial growth hormone insensitivity syndrome. Patent Number: 6,207,640. Date of Patent: March 27, 2001.

Fong S, Ferrara N, **Goddard A**, Godowski PJ, Gurney AL, Hillan K and Williams PM. Nucleic acids encoding NL-3. Patent Number: 6,074,873. Date of Patent: June 13, 2000

Attie K, Carlsson LMS, Gesundheit N and **Goddard A**. Treatment of partial growth hormone insensitivity syndrome. Patent Number: 5,824,642. Date of Patent: October 20, 1998

Attie K, Carlsson LMS, Gesundheit N and **Goddard A**. Treatment of partial growth hormone insensitivity syndrome. Patent Number: 5,646,113. Date of Patent: July 8, 1997

Multiple additional provisional applications filed



## PUBLICATIONS

Seshasayee D, Dowd P, Gu Q, Erickson S, **Goddard AD** Comparative sequence analysis of the *HER2* locus in mouse and man. Manuscript in preparation.

Abuzzahab MJ, **Goddard A**, Grigorescu F, Lautier C, Smith RJ and Chernauek SD. Human IGF-1 receptor mutations resulting in pre- and post-natal growth retardation. Manuscript in preparation.

Aggarwal S, Xie, M-H, Foster J, Frantz G, Stinson J, Corpuz RT, Simmons L, Hillan K, Yansura DG, Vandlen RL, **Goddard AD** and Gurney AL. FHFR, a novel receptor for the fibroblast growth factors. Manuscript submitted.

Adams SH, Chui C, Schilbach SL, Yu XX, **Goddard AD**, Grimaldi JC, Lee J, Dowd P, Colman S., Lewin DA. (2001) BFIT, a unique acyl-CoA thioesterase induced in thermogenic brown adipose tissue: Cloning, organization of the human gene, and assessment of a potential link to obesity. *Biochemical Journal* **360**: 135-142.

Lee J, Ho WH, Maruoka M, Corpuz RT, Baldwin DT, Foster JS, **Goddard AD**, Yansura DG, Vandlen RL, Wood WI, Gurney AL. (2001) IL-17E, a novel proinflammatory ligand for the IL-17 receptor homolog IL-17Rh1. *Journal of Biological Chemistry* **276**(2): 1660-1664.

Xie M-H, Aggarwal S, Ho W-H, Foster J, Zhang Z, Stinson J, Wood WI, **Goddard AD** and Gurney AL. (2000) Interleukin (IL)-22, a novel human cytokine that signals through the interferon-receptor related proteins CRF2-4 and IL-22R. *Journal of Biological Chemistry* **275**: 31335-31339.

Weiss GA, Watanabe CK, Zhong A, **Goddard A** and Sidhu SS. (2000) Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci. USA* **97**: 8950-8954.

Guo S, Yamaguchi Y, Schilbach S, Wada T.; Lee J, **Goddard A**, French D, Handa H, Rosenthal A. (2000) A regulator of transcriptional elongation controls vertebrate neuronal development. *Nature* **408**: 366-369.

Yan M, Wang L-C, Hymowitz SG, Schilbach S, Lee J, **Goddard A**, de Vos AM, Gao WQ, Dixit VM. (2000) Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors. *Science* **290**: 523-527.

Sehl PD, Tai JTN, Hillan KJ, Brown LA, **Goddard A**, Yang R, Jin H and Lowe DG. (2000) Application of cDNA microarrays in determining molecular phenotype in cardiac growth, development, and response to injury. *Circulation* **101**: 1990-1999.

Guo S, Brush J, Teraoka H, **Goddard A**, Wilson SW, Mullins MC and Rosenthal A. (1999) Development of noradrenergic neurons in the zebrafish hindbrain requires BMP, FGF8, and the homeodomain protein soulless/Phox2A. *Neuron* **24**: 555-566.

Stone D, Murone, M, Luoh, S, Ye W, Armanini P, Gurney A, Phillips HS, Brush, J, **Goddard A**, de Sauvage FJ and Rosenthal A. (1999) Characterization of the human suppressor of fused; a negative regulator of the zinc-finger transcription factor Gli. *J. Cell Sci.* **112**: 4437-4448.

Xie M-H, Holcomb I, Deuel B, Dowd P, Huang A, Vagts A, Foster J, Liang J, Brush J, Gu Q, Hillan K, **Goddard A** and Gurney, A.L. (1999) FGF-19, a novel fibroblast growth factor with unique specificity for FGFR4. *Cytokine* **11**: 729-735.

- Yan M, Lee J, Schilbach S, **Goddard A** and Dixit V. (1999) mE10, a novel caspase recruitment domain-containing proapoptotic molecule. *J. Biol. Chem.* **274**(15): 10287-10292.
- Gurney AL, Marsters SA, Huang RM, Pitti RM, Mark DT, Baldwin DT, Gray AM, Dowd P, Brush J, Heldens S, Schow P, **Goddard AD**, Wood WI, Baker KP, Godowski PJ and Ashkenazi A. (1999) Identification of a new member of the tumor necrosis factor family and its receptor, a human ortholog of mouse GITR. *Current Biology* **9**(4): 215-218.
- Ridgway JBB, Ng E, Kern JA, Lee J, Brush J, **Goddard A** and Carter P. (1999) Identification of a human anti-CD55 single-chain Fv by subtractive panning of a phage library using tumor and nontumor cell lines. *Cancer Research* **59**: 2718-2723.
- Pitti RM, Marsters SA, Lawrence DA, Roy M, Kischkel FC, Dowd P, Huang A, Donahue CJ, Sherwood SW, Baldwin DT, Godowski PJ, Wood WI, Gurney AL, Hillan KJ, Cohen RL, **Goddard AD**, Botstein D and Ashkenazi A. (1998) Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer. *Nature* **396**(6712): 699-703.
- Pennica D, Swanson TA, Welsh JW, Roy MA, Lawrence DA, Lee J, Brush J, Taneyhill LA, Deuel B, Lew M, Watanabe C, Cohen RL, Melhem MF, Finley GG, Quirke P, **Goddard AD**, Hillan KJ, Gurney AL, Botstein D and Levine AJ. (1998) WISP genes are members of the connective tissue growth factor family that are up-regulated in wnt-1-transformed cells and aberrantly expressed in human colon tumors. *Proc. Natl. Acad. Sci. USA.* **95**(25): 14717-14722.
- Yang RB, Mark MR, Gray A, Huang A, Xie MH, Zhang M, **Goddard A**, Wood WI, Gurney AL and Godowski PJ. (1998) Toll-like receptor-2 mediates lipopolysaccharide-induced cellular signalling. *Nature* **395**(6699): 284-288.
- Merchant AM, Zhu Z, Yuan JQ, **Goddard A**, Adams CW, Presta LG and Carter P. (1998) An efficient route to human bispecific IgG. *Nature Biotechnology* **16**(7): 677-681.
- Marsters SA, Sheridan JP, Pitti RM, Brush J, **Goddard A** and Ashkenazi A. (1998) Identification of a ligand for the death-domain-containing receptor Apo3. *Current Biology* **8**(9): 525-528.
- Xie J, Murone M, Luoh SM, Ryan A, Gu Q, Zhang C, Bonifas JM, Lam CW, Hynes M, **Goddard A**, Rosenthal A, Epstein EH Jr. and de Sauvage FJ. (1998) Activating Smoothed mutations in sporadic basal-cell carcinoma. *Nature.* **391**(6662): 90-92.
- Marsters SA, Sheridan JP, Pitti RM, Huang A, Skubatch M, Baldwin D, Yuan J, Gurney A, **Goddard AD**, Godowski P and Ashkenazi A. (1997) A novel receptor for Apo2L/TRAIL contains a truncated death domain. *Current Biology.* **7**(12): 1003-1006.
- Hynes M, Stone DM, Dowd M, Pitts-Meek S, **Goddard A**, Gurney A and Rosenthal A. (1997) Control of cell pattern in the neural tube by the zinc finger transcription factor *Gli-1*. *Neuron* **19**: 15-26.
- Sheridan JP, Marsters SA, Pitti RM, Gurney A., Skubatch M, Baldwin D, Ramakrishnan L, Gray CL, Baker K, Wood WI, **Goddard AD**, Godowski P, and Ashkenazi A. (1997) Control of TRAIL-Induced Apoptosis by a Family of Signaling and Decoy Receptors. *Science* **277** (5327): 818-821.

**Goddard AD**, Dowd P, Chernauek S, Geffner M, Gertner J, Hintz R, Hopwood N, Kaplan S, Plotnick L, Rogol A, Rosenfield R, Saenger P, Mauras N, Hershkopf R, Angulo M and Attie, K. (1997) Partial growth hormone insensitivity: The role of growth hormone receptor mutations in idiopathic short stature. *J. Pediatr.* **131**: S51-55.

Klein RD, Sherman D, Ho WH, Stone D, Bennett GL, Moffat B, Vandlen R, Simmons L, Gu Q, Hongo JA, Devaux B, Poulsen K, Armanini M, Nozaki C, Asai N, **Goddard A**, Phillips H, Henderson CE, Takahashi M and Rosenthal A. (1997) A GPI-linked protein that interacts with Ret to form a candidate neurturin receptor. *Nature.* **387**(6634): 717-21.

Stone DM, Hynes M, Armanini M, Swanson TA, Gu Q, Johnson RL, Scott MP, Pennica D, **Goddard A**, Phillips H, Noll M, Hooper JE, de Sauvage F and Rosenthal A. (1996) The tumour-suppressor gene patched encodes a candidate receptor for Sonic hedgehog. *Nature* **384**(6605): 129-34.

Marsters SA, Sheridan JP, Donahue CJ, Pitti RM, Gray CL, **Goddard AD**, Bauer KD and Ashkenazi A. (1996) Apo-3, a new member of the tumor necrosis factor receptor family, contains a death domain and activates apoptosis and NF-kappa  $\beta$ . *Current Biology* **6**(12): 1669-76.

Rothe M, Xiong J, Shu HB, Williamson K, **Goddard A** and Goeddel DV. (1996) I-TRAF is a novel TRAF-interacting protein that regulates TRAF-mediated signal transduction. *Proc. Natl. Acad. Sci. USA* **93**: 8241-8246.

Yang M, Luoh SM, **Goddard A**, Reilly D, Henzel W and Bass S. (1996) The bglX gene located at 47.8 min on the Escherichia coli chromosome encodes a periplasmic beta-glucosidase. *Microbiology* **142**: 1659-65.

**Goddard AD** and Black DM. (1996) Familial Cancer in Molecular Endocrinology of Cancer. Waxman, J. Ed. Cambridge University Press, Cambridge UK, pp.187-215.

Treanor JJS, Goodman L, de Sauvage F, Stone DM, Poulson KT, Beck CD, Gray C, Armanini MP, Pollocks RA, Hefti F, Phillips HS, **Goddard A**, Moore MW, Buj-Bello A, Davis AM, Asai N, Takahashi M, Vandlen R, Henderson CE and Rosenthal A. (1996) Characterization of a receptor for GDNF. *Nature* **382**: 80-83.

Klein RD, Gu Q, **Goddard A** and Rosenthal A. (1996) Selection for genes encoding secreted proteins and receptors. *Proc. Natl. Acad. Sci. USA* **93**: 7108-7113.

Winslow JW, Moran P, Valverde J, Shih A, Yuan JQ, Wong SC, Tsai SP, **Goddard A**, Henzel WJ, Hefti F and Caras I. (1995) Cloning of AL-1, a ligand for an Eph-related tyrosine kinase receptor involved in axon bundle formation. *Neuron* **14**: 973-981.

Bennett BD, Zeigler FC, Gu Q, Fendly B, **Goddard AD**, Gillett N and Matthews W. (1995) Molecular cloning of a ligand for the EPH-related receptor protein-tyrosine kinase Htk. *Proc. Natl. Acad. Sci. USA* **92**: 1866-1870.

Huang X, Yuang J, **Goddard A**, Foulis A, James RF, Lernmark A, Pujol-Borrell R, Rabinovitch A, Somoza N and Stewart TA. (1995) Interferon expression in the pancreases of patients with type I diabetes. *Diabetes* **44**: 658-664.

**Goddard AD**, Yuan JQ, Fairbairn L, Dexter M, Borrow J, Kozak C and Solomon E. (1995) Cloning of the murine homolog of the leukemia-associated PML gene. *Mammalian Genome* **6**: 732-737.

**Goddard AD**, Covello R, Luoh SM, Clackson T, Attie KM, Gesundheit N, Rundle AC, Wells JA, Carlsson LMTI and The Growth Hormone Insensitivity Study Group. (1995) Mutations of the growth hormone receptor in children with idiopathic short stature. *N. Engl. J. Med.* **333**: 1093-1098.

Kuo SS, Moran P, Gripp J, Armanini M, Phillips HS, **Goddard A** and Caras IW. (1994) Identification and characterization of Batk, a predominantly brain-specific non-receptor protein tyrosine kinase related to Csk. *J. Neurosci. Res.* **38**: 705-715.

Mark MR, Scadden DT, Wang Z, Gu Q, **Goddard A** and Godowski PJ. (1994) Rse, a novel receptor-type tyrosine kinase with homology to Axl/Ufo, is expressed at high levels in the brain. *Journal of Biological Chemistry* **269**: 10720-10728.

Borrow J, Shipley J, Howe K, Kiely F, **Goddard A**, Sheer D, Srivastava A, Antony AC, Fioretos T, Mitelman F and Solomon E. (1994) Molecular analysis of simple variant translocations in acute promyelocytic leukemia. *Genes Chromosomes Cancer* **9**: 234-243.

**Goddard AD** and Solomon E. (1993) Genetics of Cancer. *Adv. Hum. Genet.* **21**: 321-376.

Borrow J, **Goddard AD**, Gibbons B, Katz F, Swirsky D, Fioretos T, Dube I, Winfield DA, Kingston J, Hagemeijer A, Rees JKH, Lister AT and Solomon E. (1992) Diagnosis of acute promyelocytic leukemia by RT-PCR: Detection of *PML-RARA* and *RARA-PML* fusion transcripts. *Br. J. Haematol.* **82**: 529-540.

**Goddard AD**, Borrow J and Solomon E. (1992) A previously uncharacterized gene, PML, is fused to the retinoic acid receptor alpha gene in acute promyelocytic leukemia. *Leukemia* **6 Suppl 3**: 117S-119S.

Zhu X, Dunn JM, **Goddard AD**, Squire JA, Becker A, Phillips RA and Gallie BL. (1992) Mechanisms of loss of heterozygosity in retinoblastoma. *Cytogenet. Cell. Genet.* **59**: 248-252.

Foulkes W, **Goddard A** and Patel K. (1991) Retinoblastoma linked with Seascale [letter]. *British Med. J.* **302**: 409.

**Goddard AD**, Borrow J, Freemont PS and Solomon E. (1991) Characterization of a novel zinc finger gene disrupted by the t(15;17) in acute promyelocytic leukemia. *Science* **254**: 1371-1374.

Solomon E, Borrow J and **Goddard AD**. (1991) Chromosomal aberrations in cancer. *Science* **254**: 1153-1160.

Pajunen L, Jones TA, **Goddard A**, Sheer D, Solomon E, Pihlajaniemi T and Kivirikko KI. (1991) Regional assignment of the human gene coding for a multifunctional peptide (P4HB) acting as the  $\beta$ -subunit of prolyl-4-hydroxylase and the enzyme protein disulfide isomerase to 17q25. *Cytogenet. Cell. Genet.* **56**: 165-168.

Borrow J, Black DM, **Goddard AD**, Yagle MK, Frischauf A.-M and Solomon E. (1991) Construction and regional localization of a *NotI* linking library from human chromosome 17q. *Genomics* **10**: 477-480.

Borrow J, **Goddard AD**, Sheer D and Solomon E. (1990) Molecular analysis of acute promyelocytic leukemia breakpoint cluster region on chromosome 17. *Science* **249**: 1577-1580.

Myers JC, Jones TA, Pohjolainen E-R, Kadri AS, **Goddard AD**, Sheer D, Solomon E and Pihlajaniemi T. (1990) Molecular cloning of 5(IV) collagen and assignment of the gene to the region of the X-chromosome containing the Alport Syndrome locus. *Am. J. Hum. Genet.* **46**: 1024-1033.

Gallie BL, Squire JA, **Goddard A**, Dunn JM, Canton M, Hinton D, Zhu X and Phillips RA. (1990) Mechanisms of oncogenesis in retinoblastoma. *Lab. Invest.* **62**: 394-408.

**Goddard AD**, Phillips RA, Greger V, Passarge E, Hopping W, Gallie BL and Horsthemke B. (1990) Use of the RB1 cDNA as a diagnostic probe in retinoblastoma families. *Clinical Genetics* **37**: 117-126.

Zhu XP, Dunn JM, Phillips RA, **Goddard AD**, Paton KE, Becker A and Gallie BL. (1989) Germline, but not somatic, mutations of the RB1 gene preferentially involve the paternal allele. *Nature* **340**: 312-314.

Gallie BL, Dunn JM, **Goddard A**, Becker A and Phillips RA. (1988) Identification of mutations in the putative retinoblastoma gene. In Molecular Biology of The Eye: Genes, Vision and Ocular Disease. UCLA Symposia on Molecular and Cellular Biology, New Series, Volume 88. J. Piatigorsky, T. Shinohara and P.S. Zelenka, Eds. Alan R. Liss, Inc., New York, 1988, pp. 427-436.

**Goddard AD**, Balakier H, Canton M, Dunn J, Squire J, Reyes E, Becker A, Phillips RA and Gallie BL. (1988) Infrequent genomic rearrangement and normal expression of the putative RB1 gene in retinoblastoma tumors. *Mol. Cell. Biol.* **8**: 2082-2088.

Squire J, Dunn J, **Goddard A**, Hoffman T, Musarella M, Willard HF, Becker AJ, Gallie BL and Phillips RA. (1986) Cloning of the esterase D gene: A polymorphic gene probe closely linked to the retinoblastoma locus on chromosome 13. *Proc. Natl. Acad. Sci. USA* **83**: 6573-6577.

Squire J, **Goddard AD**, Canton M, Becker A, Phillips RA and Gallie BL (1986) Tumour induction by the retinoblastoma mutation is independent of N-myc expression. *Nature* **322**: 555-557.

**Goddard AD**, Heddle JA, Gallie BL and Phillips RA. (1985) Radiation sensitivity of fibroblasts of bilateral retinoblastoma patients as determined by micronucleus induction *in vitro*. *Mutation Research* **152**: 31-38.

## RESEARCH

## SIMULTANEOUS AMPLIFICATION AND DETECTION OF SPECIFIC DNA SEQUENCES

Russell Higuchi\*, Gavin Dollinger<sup>1</sup>, P. Sean Walsh and Robert GriffithRoche Molecular Systems, Inc., 1400 53rd St., Emeryville, CA 94608. <sup>1</sup>Chiron Corporation, 1400 53rd St., Emeryville, CA 94608. \*Corresponding author.

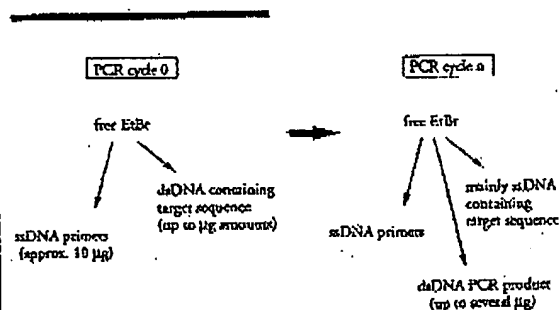
We have enhanced the polymerase chain reaction (PCR) such that specific DNA sequences can be detected without opening the reaction tube. This enhancement requires the addition of ethidium bromide (EtBr) to a PCR. Since the fluorescence of EtBr increases in the presence of double-stranded (ds) DNA an increase in fluorescence in such a PCR indicates a positive amplification, which can be easily monitored externally. In fact, amplification can be continuously monitored in order to follow its progress. The ability to simultaneously amplify specific DNA sequences and detect the product of the amplification both simplifies and improves PCR and may facilitate its automation and more widespread use in the clinic or in other situations requiring high sample throughput.

Although the potential benefits of PCR<sup>1</sup> to clinical diagnostics are well known<sup>2,3</sup>, it is still not widely used in this setting, even though it is four years since thermostable DNA polymerases<sup>4</sup> made PCR practical. Some of the reasons for its slow acceptance are high cost, lack of automation of pre- and post-PCR processing steps, and false positive results from carryover-contamination. The first two points are related in that labor is the largest contributor to cost at the present stage of PCR development. Most current assays require some form of "downstream" processing once thermocycling is done in order to determine whether the target DNA sequence was present and has amplified. These include DNA hybridization<sup>5,6</sup>, gel electrophoresis with or without use of restriction digestion<sup>7,8</sup>, HPLC<sup>9</sup>, or capillary electrophoresis<sup>10</sup>. These methods are labor-intensive, have low throughput, and are difficult to automate. The third point is also closely related to downstream processing. The handling of the PCR product in these downstream processes increases the chances that amplified DNA will spread through the typing lab, resulting in a risk of

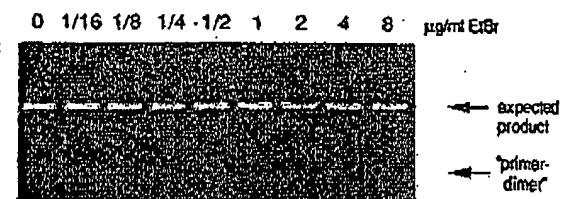
"carryover" false positives in subsequent testing<sup>11</sup>.

These downstream processing steps would be eliminated if specific amplification and detection of amplified DNA took place simultaneously within an unopened reaction vessel. Assays in which such different processes take place without the need to separate reaction components have been termed "homogeneous". No truly homogeneous PCR assay has been demonstrated to date, although progress towards this end has been reported. Chehab, et al.<sup>12</sup>, developed a PCR product detection scheme using fluorescent primers that resulted in a fluorescent PCR product. Allele-specific primers, each with different fluorescent tags, were used to indicate the genotype of the DNA. However, the unincorporated primers must still be removed in a downstream process in order to visualize the result. Recently, Holland, et al.<sup>13</sup>, developed an assay in which the endogenous 5' exonuclease assay of *Taq* DNA polymerase was exploited to cleave a labeled oligonucleotide probe. The probe would only cleave if PCR amplification had produced its complementary sequence. In order to detect the cleavage products, however, a subsequent process is again needed.

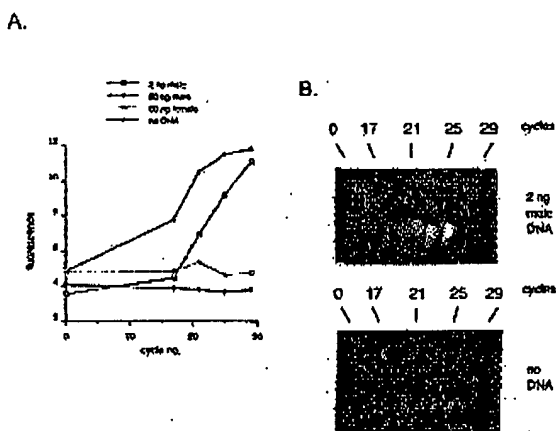
We have developed a truly homogeneous assay for PCR and PCR product detection based upon the greatly increased fluorescence that ethidium bromide and other DNA binding dyes exhibit when they are bound to dsDNA<sup>14-16</sup>. As outlined in Figure 1, a prototypic PCR



**FIGURE 1** Principle of simultaneous amplification and detection of PCR product. The components of a PCR containing EtBr that are fluorescent are listed—EtBr itself, EtBr bound to either ssDNA or dsDNA. There is a large fluorescence enhancement when EtBr is bound to DNA and binding is greatly enhanced when DNA is double-stranded. After sufficient (n) cycles of PCR, the net increase in dsDNA results in additional EtBr binding, and a net increase in total fluorescence.



**FIGURE 2** Gel electrophoresis of PCR amplification products of the human nuclear gene, HLA DQ $\alpha$ , made in the presence of increasing amounts of EtBr (up to 8  $\mu$ g/ml). The presence of EtBr has no obvious effect on the yield or specificity of amplification.



**FIGURE 3** (A) Fluorescence measurements from PCR tubes that contain 0.5  $\mu$ g/ml EtBr and that are specific for Y-chromosome repeat sequences. Five replicate PCRs were begun containing each of the DNAs specified. At each indicated cycle, one of the five replicate PCR tubes for each DNA was removed from thermocycling and its fluorescence measured. Units of fluorescence are arbitrary. (B) UV photograph of PCR tubes (0.5 ml Eppendorf-style, polypropylene micro-centrifuge tubes) containing reactions, those starting from 2 ng male DNA and control reactions without any DNA, from (A).

begins with primers that are single-stranded DNA (ssDNA), dNTPs, and DNA polymerase. An amount of dsDNA containing the target sequence (target DNA) is also typically present. This amount can vary, depending on the application, from single-cell amounts of DNA<sup>17</sup> to micrograms per PCR<sup>18</sup>. If EtBr is present, the reagents that will fluoresce, in order of increasing fluorescence, are free EtBr itself, and EtBr bound to the single-stranded DNA primers and to the double-stranded target DNA (by its intercalation between the stacked bases of the DNA double-helix). After the first denaturation cycle, target DNA will be largely single-stranded. After a PCR is completed, the most significant change is the increase in the amount of dsDNA (the PCR product itself) of up to several micrograms. Formerly free EtBr is bound to the additional dsDNA, resulting in an increase in fluorescence. There is also some decrease in the amount of ssDNA primer, but because the binding of EtBr to ssDNA is much less than to dsDNA, the effect of this change on the total fluorescence of the sample is small. The fluorescence increase can be measured by directing excitation illumination through the walls of the amplification vessel

before and after, or even continuously during, thermocycling.

## RESULTS

**PCR in the presence of EtBr.** In order to assess the effect of EtBr in PCR, amplifications of the human HLA DQ $\alpha$  gene<sup>19</sup> were performed with the dye present at concentrations from 0.06 to 8.0  $\mu$ g/ml (a typical concentration of EtBr used in staining of nucleic acids following gel electrophoresis is 0.5  $\mu$ g/ml). As shown in Figure 2, gel electrophoresis revealed little or no difference in the yield or quality of the amplification product whether EtBr was absent or present at any of these concentrations, indicating that EtBr does not inhibit PCR.

**Detection of human Y-chromosome specific sequences.** Sequence-specific, fluorescence enhancement of EtBr as a result of PCR was demonstrated in a series of amplifications containing 0.5  $\mu$ g/ml EtBr and primers specific to repeat DNA sequences found on the human Y-chromosome<sup>20</sup>. These PCRs initially contained either 60 ng male, 60 ng female, 2 ng male human or no DNA. Five replicate PCRs were begun for each DNA. After 0, 17, 21, 24 and 29 cycles of thermocycling, a PCR for each DNA was removed from the thermocycler, and its fluorescence measured in a spectrofluorometer and plotted vs. amplification cycle number (Fig. 3A). The shape of this curve reflects the fact that by the time an increase in fluorescence can be detected, the increase in DNA is becoming linear and not exponential with cycle number. As shown, the fluorescence increased about three-fold over the background fluorescence for the PCRs containing human male DNA, but did not significantly increase for negative control PCRs, which contained either no DNA or human female DNA. The more male DNA present to begin with—60 ng versus 2 ng—the fewer cycles were needed to give a detectable increase in fluorescence. Gel electrophoresis on the products of these amplifications showed that DNA fragments of the expected size were made in the male DNA containing reactions and that little DNA synthesis took place in the control samples.

In addition, the increase in fluorescence was visualized by simply laying the completed, unopened PCR tubes on a UV transilluminator and photographing them through a red filter. This is shown in figure 3B for the reactions that began with 2 ng male DNA and those with no DNA.

**Detection of specific alleles of the human  $\beta$ -globin gene.** In order to demonstrate that this approach has adequate specificity to allow genetic screening, a detection of the sickle-cell anemia mutation was performed. Figure 4 shows the fluorescence from completed amplifications containing EtBr (0.5  $\mu$ g/ml) as detected by photography of the reaction tubes on a UV transilluminator. These reactions were performed using primers specific for either the wild-type or sickle-cell mutation of the human  $\beta$ -globin gene<sup>21</sup>. The specificity for each allele is imparted by placing the sickle-mutation site at the terminal 3' nucleotide of one primer. By using an appropriate primer annealing temperature, primer extension—and thus amplification—can take place only if the 3' nucleotide of the primer is complementary to the  $\beta$ -globin allele present<sup>21,22</sup>.

Each pair of amplifications shown in Figure 4 consists of a reaction with either the wild-type allele specific (left tube) or sickle-allele specific (right tube) primers. Three different DNAs were typed: DNA from a homozygous, wild-type  $\beta$ -globin individual (AA); from a heterozygous sickle  $\beta$ -globin individual (AS); and from a homozygous sickle  $\beta$ -globin individual (SS). Each DNA (50 ng genomic DNA to start each PCR) was analyzed in triplicate (3 pairs

emocy.

ess the  
HLA  
cent at  
oncen-  
lowing  
e 2, gel  
ie yield  
Br was  
indicat.

fic se-  
nent of  
ries of  
rimers  
human  
either  
DNA.  
after 0,  
or each  
ts fluo-  
plotted  
of this  
case in  
DNA is  
umber.  
cc-fold  
ontain-  
ncrease  
her no  
DNA  
fewer  
in fluo-  
f these  
the ex-  
taining  
in the

ualized  
n a UV  
h a red  
as that  
VA.  
-globin  
sch has  
etection  
Figure  
ications  
graphy  
These  
for ci-  
human  
nparted  
ual 3'  
primer  
has am-  
c of the  
ent.<sup>21,22</sup>  
nsists of  
the (left  
Three  
zygous,  
ozygous  
ozygous  
genomic  
(3 pairs

of reactions each). The DNA type was reflected in the relative fluorescence intensities in each pair of completed amplifications. There was a significant increase in fluorescence only where a  $\beta$ -globin allele DNA matched the primer set. When measured on a spectrofluorometer (data not shown), this fluorescence was about three times that present in a PCR where both  $\beta$ -globin alleles were mismatched to the primer set. Gel electrophoresis (not shown) established that this increase in fluorescence was due to the synthesis of nearly a microgram of a DNA fragment of the expected size for  $\beta$ -globin. There was little synthesis of dsDNA in reactions in which the allele-specific primer was mismatched to both alleles.

**Continuous monitoring of a PCR.** Using a fiber optic device, it is possible to direct excitation illumination from a spectrofluorometer to a PCR undergoing thermocycling and to return its fluorescence to the spectrofluorometer. The fluorescence readout of such an arrangement, directed at an EtBr-containing amplification of Y-chromosome specific sequences from 25 ng of human male DNA, is shown in Figure 5. The readout from a control PCR with no target DNA is also shown. Thirty cycles of PCR were monitored for each.

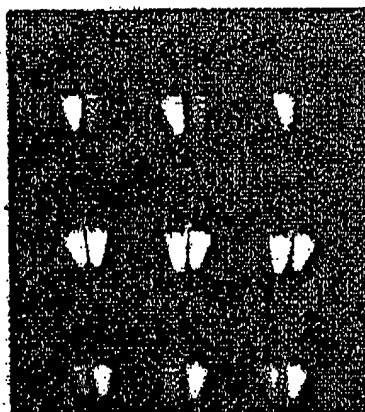
The fluorescence trace as a function of time clearly shows the effect of the thermocycling. Fluorescence intensity rises and falls inversely with temperature. The fluorescence intensity is minimum at the denaturation temperature (94°C) and maximum at the annealing/extension temperature (50°C). In the negative-control PCR, these fluorescence maxima and minima do not change significantly over the thirty thermocycles, indicating that there is little dsDNA synthesis without the appropriate target DNA, and there is little if any bleaching of EtBr during the continuous illumination of the sample.

In the PCR containing male DNA, the fluorescence maxima at the annealing/extension temperature begin to increase at about 4000 seconds of thermocycling, and continue to increase with time, indicating that dsDNA is being produced at a detectable level. Note that the fluorescence minima at the denaturation temperature do not significantly increase, presumably because at this temperature there is no dsDNA for EtBr to bind. Thus the course of the amplification is followed by tracking the fluorescence increase at the annealing temperature. Analysis of the products of these two amplifications by gel electrophoresis showed a DNA fragment of the expected size for the male DNA containing sample and no detectable DNA synthesis for the control sample.

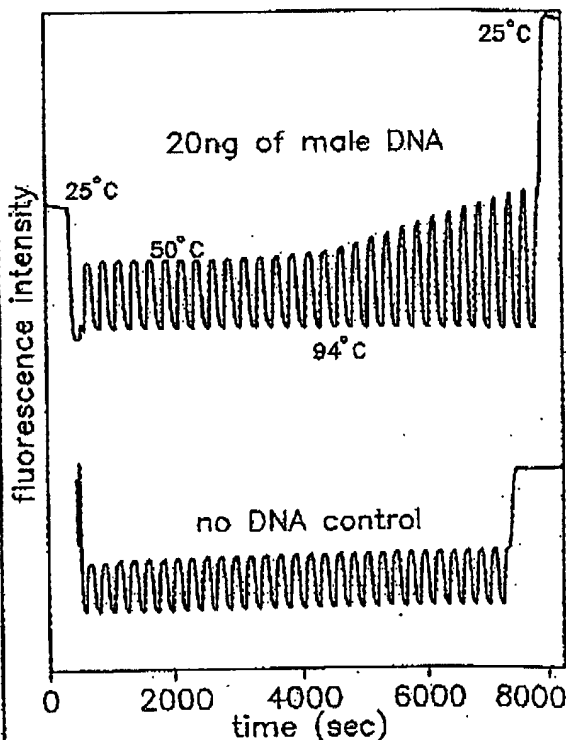
#### DISCUSSION

Downstream processes such as hybridization to a sequence-specific probe can enhance the specificity of DNA detection by PCR. The elimination of these processes means that the specificity of this homogeneous assay depends solely on that of PCR. In the case of sickle-cell disease, we have shown that PCR alone has sufficient DNA sequence specificity to permit genetic screening. Using appropriate amplification conditions, there is little non-specific production of dsDNA in the absence of the appropriate target allele.

The specificity required to detect pathogens can be more or less than that required to do genetic screening, depending on the number of pathogens in the sample and the amount of other DNA that must be taken with the sample. A difficult target is HIV, which requires detection of a viral genome that can be at the level of a few copies per thousands of host cells<sup>6</sup>. Compared with genetic screening, which is performed on cells containing at least one copy of the target sequence, HIV detection requires both more specificity and the input of more total



**FIGURE 4** UV photograph of PCR tubes containing amplifications using EtBr that are specific to wild-type (A) or sickle (S) alleles of the human  $\beta$ -globin gene. The left of each pair of tubes contains allele-specific primers to the wild-type alleles, the right tube primers to the sickle allele. The photograph was taken after 30 cycles of PCR, and the input DNAs and the alleles they contain are indicated. Fifty ng of DNA was used to begin PCR. Typing was done in triplicate (3 pairs of PCRs) for each input DNA.



**FIGURE 5** Continuous, real-time monitoring of a PCR. A fiber optic was used to carry excitation light to a PCR in progress and also emitted light back to a fluorometer (see Experimental Protocol). Amplification using human male-DNA specific primers in a PCR starting with 20 ng of human male DNA (top), or in a control PCR without DNA (bottom), were monitored. Thirty cycles of PCR were followed for each. The temperature cycled between 94°C (denaturation) and 50°C (annealing and extension). Note in the male DNA PCR, the cycle (time) dependent increase in fluorescence at the annealing/extension temperature.



DNA—up to microgram amounts—in order to have sufficient numbers of target sequences. This large amount of starting DNA in an amplification significantly increases the background fluorescence over which any additional fluorescence produced by PCR must be detected. An additional complication that occurs with targets in low copy-number is the formation of the “primer-dimer” artifact. This is the result of the extension of one primer using the other primer as a template. Although this occurs infrequently, once it occurs the extension product is a substrate for PCR amplification, and can compete with true PCR targets if those targets are rare. The primer-dimer product is of course dsDNA and thus is a potential source of false signal in this homogeneous assay.

To increase PCR specificity and reduce the effect of primer-dimer amplification, we are investigating a number of approaches, including the use of nested-primer amplifications that take place in a single tube<sup>23</sup>, and the “hot-start”, in which nonspecific amplification is reduced by raising the temperature of the reaction before DNA synthesis begins<sup>24</sup>. Preliminary results using these approaches suggest that primer-dimer is effectively reduced and it is possible to detect the increase in EtBr fluorescence in a PCR instigated by a single HIV genome in a background of  $10^5$  cells. With larger numbers of cells, the background fluorescence contributed by genomic DNA becomes problematic. To reduce this background, it may be possible to use sequence-specific DNA-binding dyes that can be made to preferentially bind PCR product over genomic DNA by incorporating the dye-binding DNA sequence into the PCR product through a 5′ “add-on” to the oligonucleotide primer<sup>24</sup>.

We have shown that the detection of fluorescence generated by an EtBr-containing PCR is straightforward, both once PCR is completed and continuously during thermocycling. The ease with which automation of specific DNA detection can be accomplished is the most promising aspect of this assay. The fluorescence analysis of completed PCRs is already possible with existing instrumentation in 96-well format<sup>25</sup>. In this format, the fluorescence in each PCR can be quantitated before, after, and even at selected points during thermocycling by moving the rack of PCRs to a 96-microwell plate fluorescence reader<sup>26</sup>.

The instrumentation necessary to continuously monitor multiple PCRs simultaneously is also simple in principle. A direct extension of the apparatus used here is to have multiple fiberoptics transmit the excitation light and fluorescent emissions to and from multiple PCRs. The ability to monitor multiple PCRs continuously may allow quantitation of target DNA copy number. Figure 3 shows that the larger the amount of starting target DNA, the sooner during PCR a fluorescence increase is detected. Preliminary experiments (Higuchi and Dollinger, manuscript in preparation) with continuous monitoring have shown a sensitivity to two-fold differences in initial target DNA concentration.

Conversely, if the number of target molecules is known—as it can be in genetic screening—continuous monitoring may provide a means of detecting false positive and false negative results. With a known number of target molecules, a true positive would exhibit detectable fluorescence by a predictable number of cycles of PCR. Increases in fluorescence detected before or after that cycle would indicate potential artifacts. False negative results due to, for example, inhibition of DNA polymerase, may be detected by including within each PCR an inefficiently amplifying marker. This marker results in a fluorescence increase only after a large number of cycles—many more than are necessary to detect a true

positive. If a sample fails to have a fluorescence increase after this many cycles, inhibition may be suspected. Since, in this assay, conclusions are drawn based on the presence or absence of fluorescence signal alone, such controls may be important. In any event, before any test based on this principle is ready for the clinic, an assessment of its false positive/false negative rates will need to be obtained using a large number of known samples.

In summary, the inclusion in PCR of dyes whose fluorescence is enhanced upon binding dsDNA makes it possible to detect specific DNA amplification from outside the PCR tube. In the future, instruments based upon this principle may facilitate the more widespread use of PCR in applications that demand the high throughput of samples.

#### EXPERIMENTAL PROTOCOL

**Human HLA-DQ $\alpha$  gene amplifications containing EtBr.** PCRs were set up in 100  $\mu$ l volumes containing 10 mM Tris-HCl, pH 8.3; 50 mM KCl; 4 mM MgCl<sub>2</sub>; 2.5 units of Taq DNA polymerase (Perkin-Elmer Cetus, Norwalk, CT); 20 pmole each of human HLA-DQ $\alpha$  gene specific oligonucleotide primers GH26 and GH27<sup>19</sup> and approximately  $10^5$  copies of DQ $\alpha$  PCR product diluted from a previous reaction. Ethidium bromide (EtBr; Sigma) was used at the concentrations indicated in Figure 2. Thermocycling proceeded for 20 cycles in a model 480 thermocycler (Perkin-Elmer Cetus, Norwalk, CT) using a “step-cycle” program of 94°C for 1 min, denaturation and 60°C for 30 sec, annealing and 72°C for 30 sec, extension.

**Y-chromosome specific PCR.** PCRs (100  $\mu$ l total reaction volume) containing 0.5  $\mu$ g/ml EtBr were prepared as described for HLA-DQ $\alpha$ , except with different primers and target DNAs. These PCRs contained 15 pmole each male DNA-specific primers Y1.1 and Y1.2<sup>20</sup>, and either 60 ng male, 60 ng female, 2 ng male, or no human DNA. Thermocycling was 94°C for 1 min, and 60°C for 1 min using a “step-cycle” program. The number of cycles for a sample were as indicated in Figure 3. Fluorescence measurement is described below.

**Allele-specific, human  $\beta$ -globin gene PCR.** Amplifications of 100  $\mu$ l volume using 0.5  $\mu$ g/ml of EtBr were prepared as described for HLA-DQ $\alpha$  above except with different primers and target DNAs. These PCRs contained either primer pair HGP2/Hp14A (wild-type globin specific primers) or HGP2/Hp14S (sickle-globin specific primers) at 10 pmole each primer per PCR. These primers were developed by Wu et al.<sup>21</sup>. Three different target DNAs were used in separate amplifications—50 ng each of human DNA that was homozygous for the sickle trait (SS), DNA that was heterozygous for the sickle trait (AS), or DNA that was homozygous for the w.t. globin (AA). Thermocycling was for 30 cycles at 94°C for 1 min, and 55°C for 1 min, using a “step-cycle” program. An annealing temperature of 55°C had been shown by Wu et al.<sup>21</sup> to provide allele-specific amplification. Completed PCRs were photographed through a red filter (Wratten 23A) after placing the reaction tubes atop a model TM-36 (transilluminator (UV-products San Gabriel, CA).

**Fluorescence measurement.** Fluorescence measurements were made on PCRs containing EtBr in a Fluorolog-2 fluorometer (SPEX, Edison, NJ). Excitation was at the 500 nm band with about 2 nm bandwidth with a GG 435 nm cut-off filter (Melles Crist, Inc., Irvine, CA) to exclude second-order light. Emitted light was detected at 570 nm with a bandwidth of about 7 nm. An OG 530 nm cut-off filter was used to remove the excitation light.

**Continuous fluorescence monitoring of PCR.** Continuous monitoring of a PCR in progress was accomplished using the spectrofluorometer and settings described above as well as a fiberoptic accessory (SPEX cat. no. 1950) to both send excitation light to, and receive emitted light from, a PCR placed in a well of a model 480 thermocycler (Perkin-Elmer Cetus). The probe end of the fiberoptic cable was attached with “5 minute-epoxy” to the open top of a PCR tube (a 0.5 ml polypropylene centrifuge tube with its cap removed) effectively sealing it. The exposed top of the PCR tube and the end of the fiberoptic cable were shielded from room light and the room lights were kept dimmed during each run. The monitored PCR was an amplification of Y-chromosome-specific repeat sequences as described above, except using an annealing/extension temperature of 50°C. The reaction was covered with mineral oil (2 drops) to prevent evaporation. Thermocycling and fluorescence measurement were started simultaneously. A time-base scan with a 10 second integration time

was used and the emission signal was ratioed to the excitation signal to control for changes in light-source intensity. Data were collected using the dm3000f, version 2.5 (SPEX) data system.

#### Acknowledgments

We thank Bob Jones for help with the spectrofluorometric measurements and Heatherbell Fong for editing this manuscript.

#### References

- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. 1986. Specific enzymatic amplification of DNA *in vitro*: The polymerase chain reaction. *CSHQB* 51:263-273.
- White, T. J., Arnheim, N. and Erlich, H. A. 1989. The polymerase chain reaction. *Trends Genet.* 5:185-189.
- Erlich, H. A., Gelfand, D. and Sninsky, J. J. 1991. Recent advances in the polymerase chain reaction. *Science* 252:1643-1651.
- Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. and Erlich, H. A. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239:487-491.
- Saiki, R. K., Walsh, P. S., Levenson, C. H. and Erlich, H. A. 1989. Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc. Natl. Acad. Sci. USA* 86:6230-6234.
- Kwok, S. Y., Mack, D. H., Mullis, K. B., Poiesz, B. J., Ehrlich, G. D., Blair, D. and Friedman-Kien, A. S. 1987. Identification of human immunodeficiency virus sequences by using *in vitro* enzymatic amplification and oligomer cleavage detection. *J. Virol.* 61:1690-1694.
- Chehab, F. F., Doherty, M., Cai, S. F., Kan, Y. W., Cooper, S. and Rubin, E. M. 1987. Detection of sickle cell anemia and thalassemia. *Nature* 329:293-294.
- Horn, G. T., Richards, B. and Klugger, K. W. 1989. Amplification of a highly polymorphic VNTR segment by the polymerase chain reaction. *Nuc. Acids Res.* 16:2140.
- Katz, E. D. and Dong, M. W. 1990. Rapid analysis and purification of polymerase chain reaction products by high-performance liquid chromatography. *Biotechniques* 8:546-555.
- Heiger, D. N., Cohen, A. S. and Karger, B. L. 1990. Separation of DNA restriction fragments by high performance capillary electrophoresis with low and zero crosslinked polyacrylamide using co-solvents and pulsed electric fields. *J. Chromatogr.* 516:33-48.
- Kwok, S. Y. and Higuchi, R. G. 1989. Avoiding false positives with PCR. *Nature* 339:237-238.
- Chehab, F. F. and Kan, Y. W. 1989. Detection of specific DNA sequences by fluorescence amplification: a color complementation assay. *Proc. Natl. Acad. Sci. USA* 86:9178-9182.
- Holland, P. M., Abramson, R. D., Watson, R. and Gelfand, D. H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5' to 3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. Natl. Acad. Sci. USA* 88:7276-7280.
- Markovits, J., Roques, B. P. and Le Pecq, J. B. 1979. Ethidium dimer: a new reagent for the fluorimetric determination of nucleic acids. *Anal. Biochem.* 94:259-264.
- Kapuscinski, J. and Szec, W. 1979. Interactions of 4',6-diamidino-2-phenylindole with synthetic polynucleotides. *Nuc. Acids Res.* 6:3519-3534.
- Searle, M. S. and Embrey, K. J. 1990. Sequence-specific interaction of Hoechst 33258 with the minor groove of an adenine-tract DNA duplex studied in solution by <sup>1</sup>H NMR spectroscopy. *Nuc. Acids Res.* 18:3755-3762.
- Li, H. H., Gyllenstein, U. B., Cui, X. F., Saiki, R. K., Erlich, H. A. and Arnheim, N. 1988. Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* 335:414-417.
- Abbott, M. A., Poiesz, B. J., Byrne, B. C., Kwok, S. Y., Sninsky, J. J. and Erlich, H. A. 1988. Enzymatic gene amplification: qualitative and quantitative methods for detecting proviral DNA amplified *in vitro*. *J. Infect. Dis.* 158:1158.
- Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B. and Erlich, H. A. 1986. Analysis of enzymatically amplified  $\beta$ -globin and HLA-DQ $\alpha$  DNA with allele-specific oligonucleotide probes. *Nature* 324:163-166.
- Kogan, S. G., Doherty, M. and Gitshier, J. 1987. An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. *N. Engl. J. Med.* 317:985-990.
- Wu, D. Y., Ugazochi, L., Pal, B. K. and Wallace, R. B. 1989. Allele-specific enzymatic amplification of  $\beta$ -globin genomic DNA for diagnosis of sickle cell anemia. *Proc. Natl. Acad. Sci. USA* 86:2757-2760.
- Kwok, S., Kellogg, D. E., McKinney, N., Spasic, D., Guda, L., Levenson, C. and Sninsky, J. J. 1990. Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nuc. Acids Res.* 18:999-1005.
- Chou, Q., Russell, M., Birch, D., Raymond, J. and Bloch, W. 1992. Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. *Submitted.*
- Higuchi, R. 1989. Using PCR to engineer DNA. p. 61-70. *In: PCR Technology*. H. A. Erlich (Ed.). Stockton Press, New York, N.Y.
- Haff, L., Atwood, J. C., DiCesare, J., Katz, E., Pionta, E., Williams, J. F. and Wondenberg, T. 1991. A high-performance system for automation of the polymerase chain reaction. *Biotechniques* 10:102-109, 106-112.
- Tumosa, N. and Kahana, L. 1989. Fluorescent EIA screening of monoclonal antibodies to cell surface antigens. *J. Immun. Meth.* 116:59-63.

# IBL

IMMUNO BIOLOGICAL LABORATORIES

## sCD-14 ELISA

## Trauma, Shock and Sepsis

The CD-14 molecule is expressed on the surface of monocytes and some macrophages. Membrane-bound CD-14 is a receptor for lipopolysaccharide (LPS) complexed to LPS-Binding-Protein (LBP). The concentration of its soluble form is altered under certain pathological conditions. There is evidence for an important role of sCD-14 with polytrauma, sepsis, burnings and inflammations. During septic conditions and acute infections it seems to be a prognostic marker and is therefore of value in monitoring these patients.

IBL offers an ELISA for quantitative determination of soluble CD-14 in human serum, -plasma, cell-culture supernatants and other biological fluids.

Assay features: 12x8 determinations (microtiter strips),  
precoated with a specific monoclonal antibody,  
2x1 hour incubation,  
standard range: 3 - 96 ng/ml  
detection limit: 1 ng/ml  
CV: intra- and interassay < 8%

For more information call or fax

GESELLSCHAFT FÜR IMMUNCHEMIE UND -BIOLOGIE MBH  
OSTERSTRASSE 86 · D-2000 HAMBURG 20 · GERMANY · TEL. +40/491 00 61-64 · FAX +40/40 11 98

BIO TECHNOLOGY VOL 10 APRIL 1992

417

**GENENTECH, INC.**  
1 DNA Way  
South San Francisco, CA 94080 USA  
Phone: (650) 225-1000

---

FAX: (650) 952-9881

---

**FACSIMILE TRANSMITTAL**

---

**Date:** 19 July 2004

**To:** Anna Barry  
Heller Ehrman

**Re:** Higuchi reference

**Fax No:** 324-6638

**From:** Patty Tobin, Assistant to Elizabeth M. Barnes, Ph.D.  
Genentech, Inc. Legal Department

**Number of Pages including this cover sheet:** 6

## RESEARCH

## SIMULTANEOUS AMPLIFICATION AND DETECTION OF SPECIFIC DNA SEQUENCES

Russell Higuchi\*, Gavin Dollinger<sup>1</sup>, P. Sean Walsh and Robert GriffithRoche Molecular Systems, Inc., 1400 53rd St., Emeryville, CA 94608. <sup>1</sup>Chiron Corporation, 1400 53rd St., Emeryville, CA 94608. \*Corresponding author.

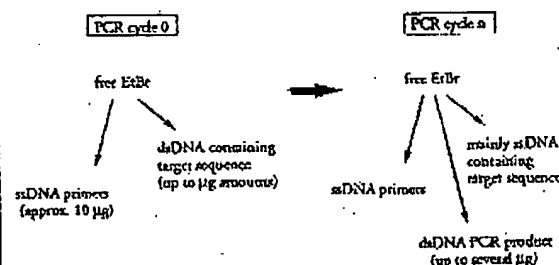
We have enhanced the polymerase chain reaction (PCR) such that specific DNA sequences can be detected without opening the reaction tube. This enhancement requires the addition of ethidium bromide (EtBr) to a PCR. Since the fluorescence of EtBr increases in the presence of double-stranded (ds) DNA an increase in fluorescence in such a PCR indicates a positive amplification, which can be easily monitored externally. In fact, amplification can be continuously monitored in order to follow its progress. The ability to simultaneously amplify specific DNA sequences and detect the product of the amplification both simplifies and improves PCR and may facilitate its automation and more widespread use in the clinic or in other situations requiring high sample throughput.

Although the potential benefits of PCR<sup>1</sup> to clinical diagnostics are well known<sup>2,3</sup>, it is still not widely used in this setting, even though it is four years since thermostable DNA polymerases<sup>4</sup> made PCR practical. Some of the reasons for its slow acceptance are high cost, lack of automation of pre- and post-PCR processing steps, and false positive results from carryover-contamination. The first two points are related in that labor is the largest contributor to cost at the present stage of PCR development. Most current assays require some form of "downstream" processing once thermocycling is done in order to determine whether the target DNA sequence was present and has amplified. These include DNA hybridization<sup>5,6</sup>, gel electrophoresis with or without use of restriction digestion<sup>7,8</sup>, HPLC<sup>9</sup>, or capillary electrophoresis<sup>10</sup>. These methods are labor-intensive, have low throughput, and are difficult to automate. The third point is also closely related to downstream processing. The handling of the PCR product in these downstream processes increases the chances that amplified DNA will spread through the typing lab, resulting in a risk of

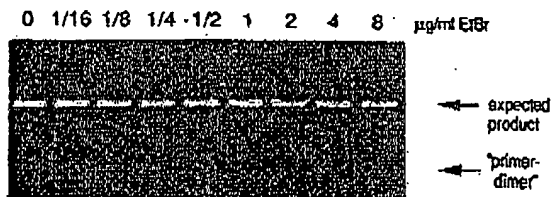
"carryover" false positives in subsequent testing<sup>11</sup>.

These downstream processing steps would be eliminated if specific amplification and detection of amplified DNA took place simultaneously within an unopened reaction vessel. Assays in which such different processes take place without the need to separate reaction components have been termed "homogeneous". No truly homogeneous PCR assay has been demonstrated to date, although progress towards this end has been reported. Chehab, et al.<sup>12</sup>, developed a PCR product detection scheme using fluorescent primers that resulted in a fluorescent PCR product. Allele-specific primers, each with different fluorescent tags, were used to indicate the genotype of the DNA. However, the unincorporated primers must still be removed in a downstream process in order to visualize the result. Recently, Holland, et al.<sup>13</sup>, developed an assay in which the endogenous 5' exonuclease assay of *Taq* DNA polymerase was exploited to cleave a labeled oligonucleotide probe. The probe would only cleave if PCR amplification had produced its complementary sequence. In order to detect the cleavage products, however, a subsequent process is again needed.

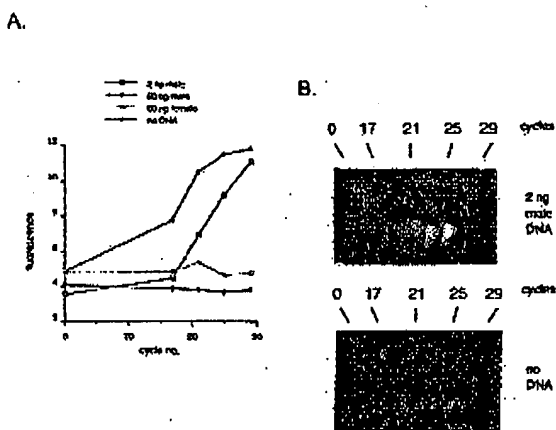
We have developed a truly homogeneous assay for PCR and PCR product detection based upon the greatly increased fluorescence that ethidium bromide and other DNA binding dyes exhibit when they are bound to ds-DNA<sup>14-16</sup>. As outlined in Figure 1, a prototypic PCR



**FIGURE 1** Principle of simultaneous amplification and detection of PCR product. The components of a PCR containing EtBr that are fluorescent are listed—EtBr itself, EtBr bound to either ssDNA or dsDNA. There is a large fluorescence enhancement when EtBr is bound to DNA and binding is greatly enhanced when DNA is double-stranded. After sufficient (n) cycles of PCR, the net increase in dsDNA results in additional EtBr binding, and a net increase in total fluorescence.



**FIGURE 2** Gel electrophoresis of PCR amplification products of the human, nuclear gene, HLA DQ $\alpha$ , made in the presence of increasing amounts of EtBr (up to 8  $\mu$ g/ml). The presence of EtBr has no obvious effect on the yield or specificity of amplification.



**FIGURE 3** (A) Fluorescence measurements from PCRs that contain 0.5  $\mu$ g/ml EtBr and that are specific for Y-chromosome repeat sequences. Five replicate PCRs were begun containing each of the DNAs specified. At each indicated cycle, one of the five replicate PCRs for each DNA was removed from thermocycling and its fluorescence measured. Units of fluorescence are arbitrary. (B) UV photograph of PCR tubes (0.5 ml Eppendorf-style, polypropylene micro-centrifuge tubes) containing reactions, those starting from 2 ng male DNA and control reactions without any DNA, from (A).

begins with primers that are single-stranded DNA (ssDNA), dNTPs, and DNA polymerase. An amount of dsDNA containing the target sequence (target DNA) is also typically present. This amount can vary, depending on the application, from single-cell amounts of DNA<sup>17</sup> to micrograms per PCR<sup>18</sup>. If EtBr is present, the reagents that will fluoresce, in order of increasing fluorescence, are free EtBr itself, and EtBr bound to the single-stranded DNA primers and to the double-stranded target DNA (by its intercalation between the stacked bases of the DNA double-helix). After the first denaturation cycle, target DNA will be largely single-stranded. After a PCR is completed, the most significant change is the increase in the amount of dsDNA (the PCR product itself) of up to several micrograms. Formerly free EtBr is bound to the additional dsDNA, resulting in an increase in fluorescence. There is also some decrease in the amount of ssDNA primer, but because the binding of EtBr to ssDNA is much less than to dsDNA, the effect of this change on the total fluorescence of the sample is small. The fluorescence increase can be measured by directing excitation illumination through the walls of the amplification vessel

before and after, or even continuously during, thermocycling.

## RESULTS

**PCR in the presence of EtBr.** In order to assess the effect of EtBr in PCR, amplifications of the human HLA DQ $\alpha$  gene<sup>19</sup> were performed with the dye present at concentrations from 0.06 to 8.0  $\mu$ g/ml (a typical concentration of EtBr used in staining of nucleic acids following gel electrophoresis is 0.5  $\mu$ g/ml). As shown in Figure 2, gel electrophoresis revealed little or no difference in the yield or quality of the amplification product whether EtBr was absent or present at any of these concentrations, indicating that EtBr does not inhibit PCR.

**Detection of human Y-chromosome specific sequences.** Sequence-specific, fluorescence enhancement of EtBr as a result of PCR was demonstrated in a series of amplifications containing 0.5  $\mu$ g/ml EtBr and primers specific to repeat DNA sequences found on the human Y-chromosome<sup>20</sup>. These PCRs initially contained either 60 ng male, 60 ng female, 2 ng male human or no DNA. Five replicate PCRs were begun for each DNA. After 0, 17, 21, 24 and 29 cycles of thermocycling, a PCR for each DNA was removed from the thermocycler, and its fluorescence measured in a spectrofluorometer and plotted vs. amplification cycle number (Fig. 3A). The shape of this curve reflects the fact that by the time an increase in fluorescence can be detected, the increase in DNA is becoming linear and not exponential with cycle number. As shown, the fluorescence increased about three-fold over the background fluorescence for the PCRs containing human male DNA, but did not significantly increase for negative control PCRs, which contained either no DNA or human female DNA. The more male DNA present to begin with—60 ng versus 2 ng—the fewer cycles were needed to give a detectable increase in fluorescence. Gel electrophoresis on the products of these amplifications showed that DNA fragments of the expected size were made in the male DNA containing reactions and that little DNA synthesis took place in the control samples.

In addition, the increase in fluorescence was visualized by simply laying the completed, unopened PCRs on a UV transilluminator and photographing them through a red filter. This is shown in figure 3B for the reactions that began with 2 ng male DNA and those with no DNA.

**Detection of specific alleles of the human  $\beta$ -globin gene.** In order to demonstrate that this approach has adequate specificity to allow genetic screening, a detection of the sickle-cell anemia mutation was performed. Figure 4 shows the fluorescence from completed amplifications containing EtBr (0.5  $\mu$ g/ml) as detected by photography of the reaction tubes on a UV transilluminator. These reactions were performed using primers specific for either the wild-type or sickle-cell mutation of the human  $\beta$ -globin gene<sup>21</sup>. The specificity for each allele is imparted by placing the sickle-mutation site at the terminal 3' nucleotide of one primer. By using an appropriate primer annealing temperature, primer extension—and thus amplification—can take place only if the 3' nucleotide of the primer is complementary to the  $\beta$ -globin allele present<sup>21,22</sup>.

Each pair of amplifications shown in Figure 4 consists of a reaction with either the wild-type allele specific (left tube) or sickle-allele specific (right tube) primers. Three different DNAs were typed: DNA from a homozygous, wild-type  $\beta$ -globin individual (AA); from a heterozygous sickle  $\beta$ -globin individual (AS); and from a homozygous sickle  $\beta$ -globin individual (SS). Each DNA (50 ng genomic DNA to start each PCR) was analyzed in triplicate (3 pairs

emocy.

ess the  
HLA  
tent at  
oncen-  
lowing  
e 2, gel  
ie yield  
Br was  
indicat.

Se se-  
nent of  
ries of  
rimers  
human  
either  
DNA.  
fter 0,  
or each  
ts fluo-  
plotted  
of this  
case in  
DNA is  
umber.  
ec-fold  
ontain-  
ncrease  
her no  
DNA  
fewer  
in fluo-  
f these  
the ex-  
taining  
in the

ualized  
n a UV  
h a red  
ons that  
VA.  
-globin  
ich has  
etection  
Figure  
ications  
graphy  
These  
for ci-  
human  
nparted  
ual 3'  
primer  
has am-  
c of the  
ent<sup>11,22</sup>  
nsists of  
the (left  
Three  
zygous  
ozygous  
zygous  
zygous  
(3 pairs

of reactions each). The DNA type was reflected in the relative fluorescence intensities in each pair of completed amplifications. There was a significant increase in fluorescence only where a  $\beta$ -globin allele matched the primer set. When measured on a spectrofluorometer (data not shown), this fluorescence was about three times that present in a PCR where both  $\beta$ -globin alleles were mismatched to the primer set. Gel electrophoresis (not shown) established that this increase in fluorescence was due to the synthesis of nearly a microgram of a DNA fragment of the expected size for  $\beta$ -globin. There was little synthesis of dsDNA in reactions in which the allele-specific primer was mismatched to both alleles.

**Continuous monitoring of a PCR.** Using a fiber optic device, it is possible to direct excitation illumination from a spectrofluorometer to a PCR undergoing thermocycling and to return its fluorescence to the spectrofluorometer. The fluorescence readout of such an arrangement, directed at an EtBr-containing amplification of Y-chromosome specific sequences from 25 ng of human male DNA, is shown in Figure 5. The readout from a control PCR with no target DNA is also shown. Thirty cycles of PCR were monitored for each.

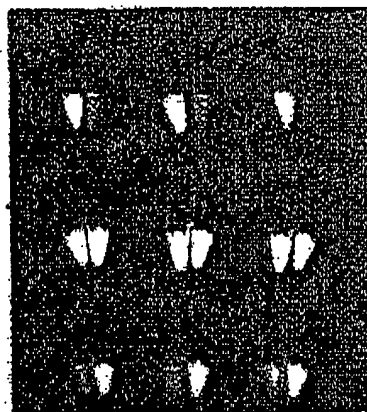
The fluorescence trace as a function of time clearly shows the effect of the thermocycling. Fluorescence intensity rises and falls inversely with temperature. The fluorescence intensity is minimum at the denaturation temperature (94°C) and maximum at the annealing/extension temperature (50°C). In the negative-control PCR, these fluorescence maxima and minima do not change significantly over the thirty thermocycles, indicating that there is little dsDNA synthesis without the appropriate target DNA, and there is little if any bleaching of EtBr during the continuous illumination of the sample.

In the PCR containing male DNA, the fluorescence maxima at the annealing/extension temperature begin to increase at about 4000 seconds of thermocycling, and continue to increase with time, indicating that dsDNA is being produced at a detectable level. Note that the fluorescence minima at the denaturation temperature do not significantly increase, presumably because at this temperature there is no dsDNA for EtBr to bind. Thus the course of the amplification is followed by tracking the fluorescence increase at the annealing temperature. Analysis of the products of these two amplifications by gel electrophoresis showed a DNA fragment of the expected size for the male DNA containing sample and no detectable DNA synthesis for the control sample.

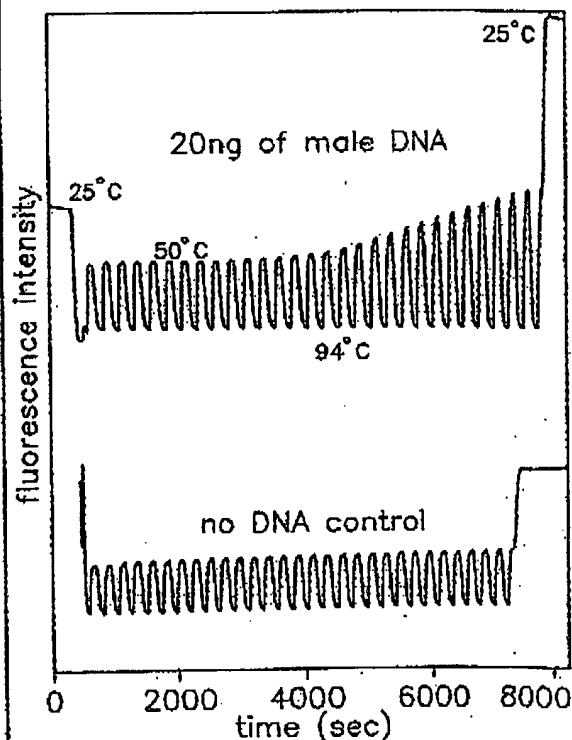
## DISCUSSION

Downstream processes such as hybridization to a sequence-specific probe can enhance the specificity of DNA detection by PCR. The elimination of these processes means that the specificity of this homogeneous assay depends solely on that of PCR. In the case of sickle-cell disease, we have shown that PCR alone has sufficient DNA sequence specificity to permit genetic screening. Using appropriate amplification conditions, there is little non-specific production of dsDNA in the absence of the appropriate target allele.

The specificity required to detect pathogens can be more or less than that required to do genetic screening, depending on the number of pathogens in the sample and the amount of other DNA that must be taken with the sample. A difficult target is HIV, which requires detection of a viral genome that can be at the level of a few copies per thousands of host cells<sup>6</sup>. Compared with genetic screening, which is performed on cells containing at least one copy of the target sequence, HIV detection requires both more specificity and the input of more total



**FIGURE 4** UV photograph of PCR tubes containing amplifications using EtBr that are specific to wild-type (A) or sickle (S) alleles of the human  $\beta$ -globin gene. The left of each pair of tubes contains allele-specific primers to the wild-type alleles, the right tube primers to the sickle allele. The photograph was taken after 30 cycles of PCR, and the input DNAs and the alleles they contain are indicated. Fifty ng of DNA was used to begin PCR. Typing was done in triplicate (3 pairs of PCRs) for each input DNA.



**FIGURE 5** Continuous, real-time monitoring of a PCR. A fiber optic was used to carry excitation light to a PCR in progress and also emitted light back to a fluorometer (see Experimental Protocol). Amplification using human male-DNA specific primers in a PCR starting with 20 ng of human male DNA (top), or in a control PCR without DNA (bottom), were monitored. Thirty cycles of PCR were followed for each. The temperature cycled between 94°C (denaturation) and 50°C (annealing and extension). Note in the male DNA PCR, the cycle (time) dependent increase in fluorescence at the annealing/extension temperature.

DNA—up to microgram amounts—in order to have sufficient numbers of target sequences. This large amount of starting DNA in an amplification significantly increases the background fluorescence over which any additional fluorescence produced by PCR must be detected. An additional complication that occurs with targets in low copy-number is the formation of the “primer-dimer” artifact. This is the result of the extension of one primer using the other primer as a template. Although this occurs infrequently, once it occurs the extension product is a substrate for PCR amplification, and can compete with true PCR targets if those targets are rare. The primer-dimer product is of course dsDNA and thus is a potential source of false signal in this homogeneous assay.

To increase PCR specificity and reduce the effect of primer-dimer amplification, we are investigating a number of approaches, including the use of nested-primer amplifications that take place in a single tube<sup>23</sup>, and the “hot-start”, in which nonspecific amplification is reduced by raising the temperature of the reaction before DNA synthesis begins<sup>24</sup>. Preliminary results using these approaches suggest that primer-dimer is effectively reduced and it is possible to detect the increase in EtBr fluorescence in a PCR instigated by a single HIV genome in a background of  $10^5$  cells. With larger numbers of cells, the background fluorescence contributed by genomic DNA becomes problematic. To reduce this background, it may be possible to use sequence-specific DNA-binding dyes that can be made to preferentially bind PCR product over genomic DNA by incorporating the dye-binding DNA sequence into the PCR product through a 5' “add-on” to the oligonucleotide primer<sup>24</sup>.

We have shown that the detection of fluorescence generated by an EtBr-containing PCR is straightforward, both once PCR is completed and continuously during thermocycling. The ease with which automation of specific DNA detection can be accomplished is the most promising aspect of this assay. The fluorescence analysis of completed PCRs is already possible with existing instrumentation in 96-well format<sup>25</sup>. In this format, the fluorescence in each PCR can be quantitated before, after, and even at selected points during thermocycling by moving the rack of PCRs to a 96-microwell plate fluorescence reader<sup>26</sup>.

The instrumentation necessary to continuously monitor multiple PCRs simultaneously is also simple in principle. A direct extension of the apparatus used here is to have multiple fiberoptics transmit the excitation light and fluorescent emissions to and from multiple PCRs. The ability to monitor multiple PCRs continuously may allow quantitation of target DNA copy number. Figure 3 shows that the larger the amount of starting target DNA, the sooner during PCR a fluorescence increase is detected. Preliminary experiments (Higuchi and Dollinger, manuscript in preparation) with continuous monitoring have shown a sensitivity to two-fold differences in initial target DNA concentration.

Conversely, if the number of target molecules is known—as it can be in genetic screening—continuous monitoring may provide a means of detecting false positive and false negative results. With a known number of target molecules, a true positive would exhibit detectable fluorescence by a predictable number of cycles of PCR. Increases in fluorescence detected before or after that cycle would indicate potential artifacts. False negative results due to, for example, inhibition of DNA polymerase, may be detected by including within each PCR an inefficiently amplifying marker. This marker results in a fluorescence increase only after a large number of cycles—many more than are necessary to detect a true

positive. If a sample fails to have a fluorescence increase after this many cycles, inhibition may be suspected. Since, in this assay, conclusions are drawn based on the presence or absence of fluorescence signal alone, such controls may be important. In any event, before any test based on this principle is ready for the clinic, an assessment of its false positive/false negative rates will need to be obtained using a large number of known samples.

In summary, the inclusion in PCR of dyes whose fluorescence is enhanced upon binding dsDNA makes it possible to detect specific DNA amplification from outside the PCR tube. In the future, instruments based upon this principle may facilitate the more widespread use of PCR in applications that demand the high throughput of samples.

#### EXPERIMENTAL PROTOCOL

**Human HLA-DQ $\alpha$  gene amplifications containing EtBr.** PCRs were set up in 100  $\mu$ l volumes containing 10 mM Tris-HCl, pH 8.3; 50 mM KCl; 4 mM MgCl<sub>2</sub>; 2.5 units of *Taq* DNA polymerase (Perkin-Elmer Cetus, Norwalk, CT); 20 pmole each of human HLA-DQ $\alpha$  gene specific oligonucleotide primers GH26 and GH27<sup>19</sup> and approximately  $10^5$  copies of DQ $\alpha$  PCR product diluted from a previous reaction. Ethidium bromide (EtBr; Sigma) was used at the concentrations indicated in Figure 2. Thermocycling proceeded for 20 cycles in a model 480 thermocycler (Perkin-Elmer Cetus, Norwalk, CT) using a “step-cycle” program of 94°C for 1 min, denaturation and 60°C for 30 sec, annealing and 72°C for 30 sec, extension.

**Y-chromosome specific PCR.** PCRs (100  $\mu$ l total reaction volume) containing 0.5  $\mu$ g/ $\mu$ l EtBr were prepared as described for HLA-DQ $\alpha$ , except with different primers and target DNAs. These PCRs contained 15 pmole each male DNA-specific primers Y1.1 and Y1.2<sup>20</sup>, and either 60 ng male, 60 ng female, 2 ng male, or no human DNA. Thermocycling was 94°C for 1 min, and 60°C for 1 min using a “step-cycle” program. The number of cycles for a sample were as indicated in Figure 3. Fluorescence measurement is described below.

**Allele-specific, human  $\beta$ -globin gene PCR.** Amplifications of 100  $\mu$ l volume using 0.5  $\mu$ g/ $\mu$ l EtBr were prepared as described for HLA-DQ $\alpha$  above except with different primers and target DNAs. These PCRs contained either primer pair HGP2/Hp14A (wild-type globin specific primers) or HGP2/Hp14S (sickle-globin specific primers) at 10 pmole each primer per PCR. These primers were developed by Wu et al.<sup>21</sup>. Three different target DNAs were used in separate amplifications—50 ng each of human DNA that was homozygous for the sickle trait (SS), DNA that was heterozygous for the sickle trait (AS), or DNA that was homozygous for the w.t. globin (AA). Thermocycling was for 30 cycles at 94°C for 1 min, and 55°C for 1 min, using a “step-cycle” program. An annealing temperature of 55°C had been shown by Wu et al.<sup>21</sup> to provide allele-specific amplification. Completed PCRs were photographed through a red filter (Wratten 23A) after placing the reaction tubes atop a model TM-96 transilluminator (UV-products San Gabriel, CA).

**Fluorescence measurement.** Fluorescence measurements were made on PCRs containing EtBr in a Fluorolog-2 fluorometer (SPEX, Edison, NJ). Excitation was at the 500 nm band with about 2 nm bandwidth with a GG 435 nm cut-off filter (Melles Griest, Inc., Irvine, CA) to exclude second-order light. Emitted light was detected at 570 nm with a bandwidth of about 7 nm. An OG 530 nm cut-off filter was used to remove the excitation light.

**Continuous fluorescence monitoring of PCR.** Continuous monitoring of a PCR in progress was accomplished using the spectrofluorometer and settings described above as well as a fiberoptic accessory (SPEX cat. no. 1950) to both send excitation light to, and receive emitted light from, a PCR placed in a well of a model 480 thermocycler (Perkin-Elmer Cetus). The probe end of the fiberoptic cable was attached with “5 minute-epoxy” to the open top of a PCR tube (a 0.5 ml polypropylene centrifuge tube with its cap removed) effectively sealing it. The exposed top of the PCR tube and the end of the fiberoptic cable were shielded from room light and the room lights were kept dimmed during each run. The monitored PCR was an amplification of Y-chromosome-specific repeat sequences as described above, except using an annealing/extension temperature of 50°C. The reaction was covered with mineral oil (2 drops) to prevent evaporation. Thermocycling and fluorescence measurement were started simultaneously. A time-base scan with a 10 second integration time



was used and the emission signal was ratioed to the excitation signal to control for changes in light-source intensity. Data were collected using the dm3000f, version 2.5 (SPEX) data system.

#### Acknowledgments

We thank Bob Jones for help with the spectrofluorometric measurements and Heatherbell Fong for editing this manuscript.

#### References

- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. 1986. Specific enzymatic amplification of DNA *in vitro*: The polymerase chain reaction. *CSHSQB* 51:263-273.
- White, T. J., Arnheim, N. and Erlich, H. A. 1989. The polymerase chain reaction. *Trends Genet.* 5:185-189.
- Erlich, H. A., Gelfand, D. and Sninsky, J. J. 1991. Recent advances in the polymerase chain reaction. *Science* 252:1643-1651.
- Saiki, R. K., Gelfand, D. M., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. and Erlich, H. A. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239:487-491.
- Saiki, R. K., Walsh, P. S., Levenson, C. H. and Erlich, H. A. 1989. Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc. Natl. Acad. Sci. USA* 86:6230-6234.
- Kwok, S. Y., Mack, D. H., Mullis, K. B., Poiesz, B. J., Ehrlich, G. D., Blair, D. and Friedman-Kien, A. S. 1987. Identification of human immunodeficiency virus sequences by using *in vitro* enzymatic amplification and oligomer cleavage detection. *J. Virol.* 61:1690-1694.
- Chehab, F. F., Doherty, M., Cai, S. P., Kan, Y. W., Cooper, S. and Rubin, E. M. 1987. Detection of sickle cell anemia and thalassemia. *Nature* 329:293-294.
- Horn, G. T., Richards, B. and Klingler, K. W. 1989. Amplification of a highly polymorphic VNTR segment by the polymerase chain reaction. *Nuc. Acids Res.* 16:2140.
- Katz, E. D. and Dong, M. W. 1990. Rapid analysis and purification of polymerase chain reaction products by high-performance liquid chromatography. *Biotechniques* 8:546-555.
- Helger, D. N., Cohen, A. S. and Karger, B. L. 1990. Separation of DNA restriction fragments by high performance capillary electrophoresis with low and zero crosslinked polyacrylamide using continuous and pulsed electric fields. *J. Chromatogr.* 516:33-48.
- Kwok, S. Y. and Higuchi, R. G. 1989. Avoiding false positives with PCR. *Nature* 339:237-238.
- Chehab, F. F. and Kan, Y. W. 1989. Detection of specific DNA sequences by fluorescence amplification: a color complementation assay. *Proc. Natl. Acad. Sci. USA* 86:9178-9182.
- Holland, P. M., Abramson, R. D., Watson, R. and Gelfand, D. H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5' to 3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. Natl. Acad. Sci. USA* 88:7276-7280.
- Markovits, J., Roques, B. F. and Le Pecq, J. B. 1979. Ethidium dimer: a new reagent for the fluorimetric determination of nucleic acids. *Anal. Biochem.* 94:259-264.
- Kapuscinski, J. and Socr, W. 1979. Interactions of 4',6-diamidino-2-phenylindole with synthetic polynucleotides. *Nuc. Acids Res.* 6:5519-5534.
- Scarle, M. S. and Embrey, K. J. 1990. Sequence-specific interaction of Hoechst 33258 with the minor groove of an adenine-tract DNA duplex studied in solution by <sup>1</sup>H NMR spectroscopy. *Nuc. Acids Res.* 18:3753-3762.
- Li, H. H., Gyllenstein, U. B., Cui, X. F., Saiki, R. K., Erlich, H. A. and Arnheim, N. 1988. Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* 336:414-417.
- Abbott, M. A., Poiesz, B. J., Byrne, B. C., Kwok, S. Y., Sninsky, J. J. and Erlich, H. A. 1988. Enzymatic gene amplification: qualitative and quantitative methods for detecting proviral DNA amplified *in vitro*. *J. Infect. Dis.* 158:1158.
- Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B. and Erlich, H. A. 1986. Analysis of enzymatically amplified  $\beta$ -globin and HLA-DQA DNA with allele-specific oligonucleotide probes. *Nature* 324:163-166.
- Kogan, S. G., Doherty, M. and Giachieri, J. 1987. An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. *N. Engl. J. Med.* 317:988-990.
- Wu, D. Y., Uguzoglu, I., Pal, B. K. and Wallace, R. B. 1989. Allele-specific enzymatic amplification of  $\beta$ -globin genomic DNA for diagnosis of sickle cell anemia. *Proc. Natl. Acad. Sci. USA* 86:2757-2760.
- Kwok, S., Kellogg, D. E., McKinney, N., Spasic, D., Goda, L., Levenson, C. and Sninsky, J. J. 1990. Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nuc. Acids Res.* 18:999-1005.
- Chou, Q., Russell, M., Birch, D., Raymond, J. and Bloch, W. 1992. Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. Submitted.
- Higuchi, R. 1989. Using PCR to engineer DNA. p. 61-70. In: *PCR Technology*. H. A. Erlich (Ed.). Stockton Press, New York, N.Y.
- Hall, L., Atwood, J. G., DiCesare, J., Katz, E., Pionza, E., Williams, J. F. and Wondenberg, T. 1991. A high-performance system for automation of the polymerase chain reaction. *Biotechniques* 10:102-103, 106-112.
- Turner, N. and Kalwa, L. 1989. Fluorescent ELISA screening of monoclonal antibodies to cell surface antigens. *J. Immun. Med.* 116:59-63.

# IBL

IMMUNO BIOLOGICAL LABORATORIES

## sCD-14 ELISA

### Trauma, Shock and Sepsis

The CD-14 molecule is expressed on the surface of monocytes and some macrophages. Membrane-bound CD-14 is a receptor for lipopolysaccharide (LPS) complexed to LPS-Binding-Protein (LBP). The concentration of its soluble form is altered under certain pathological conditions. There is evidence for an important role of sCD-14 with polytrauma, sepsis, burnings and inflammations.

During septic conditions and acute infections it seems to be a prognostic marker and is therefore of value in monitoring these patients.

IBL offers an ELISA for quantitative determination of soluble CD-14 in human serum, -plasma, cell-culture supernatants and other biological fluids.

Assay features: 12 x 8 determinations (microtiter strips),  
precoated with a specific monoclonal antibody,  
2x1 hour incubation,  
standard range: 3 - 96 ng/ml  
detection limit: 1 ng/ml  
CV: intra- and interassay < 8%

For more information call or fax

GESELLSCHAFT FÜR IMMUNCHEMIE UND -BIOLOGIE MBH  
OSTERSTRASSE 86 · D-2000 HAMBURG 20 · GERMANY · TEL. +40/491 00 61-64 · FAX +40/40 11 98

BIOTECHNOLOGY VOL 10 APRIL 1992

417



# Oligonucleotides with Fluorescent Dyes at Opposite Ends Provide a Quenched Probe System Useful for Detecting PCR Product and Nucleic Acid Hybridization

Kenneth J. Livak, Susan J.A. Flood, Jeffrey Marmaro, William Giusti, and Karin Deetz

Perkin-Elmer, Applied Biosystems Division, Foster City, California 94404

The 5' nuclease PCR assay detects the accumulation of specific PCR product by hybridization and cleavage of a double-labeled fluorogenic probe during the amplification reaction. The probe is an oligonucleotide with both a reporter fluorescent dye and a quencher dye attached. An increase in reporter fluorescence intensity indicates that the probe has hybridized to the target PCR product and has been cleaved by the 5'→3' nucleolytic activity of *Taq* DNA polymerase. In this study, probes with the quencher dye attached to an internal nucleotide were compared with probes with the quencher dye attached to the 3'-end nucleotide. In all cases, the reporter dye was attached to the 5' end. All intact probes showed quenching of the reporter fluorescence. In general, probes with the quencher dye attached to the 3'-end nucleotide exhibited a larger signal in the 5' nuclease PCR assay than the internally labeled probes. It is proposed that the larger signal is caused by increased likelihood of cleavage by *Taq* DNA polymerase when the probe is hybridized to a template strand during PCR. Probes with the quencher dye attached to the 3'-end nucleotide also exhibited an increase in reporter fluorescence intensity when hybridized to a complementary strand. Thus, oligonucleotides with reporter and quencher dyes attached at opposite ends can be used as homogeneous hybridization probes.

A homogeneous assay for detecting the accumulation of specific PCR product that uses a double-labeled fluorogenic probe was described by Lee et al.<sup>(1)</sup> The assay exploits the 5'→3' nucleolytic activity of *Taq* DNA polymerase<sup>(2,3)</sup> and is diagramed in Figure 1. The fluorogenic probe consists of an oligonucleotide with a reporter fluorescent dye, such as a fluorescein, attached to the 5' end; and a quencher dye, such as a rhodamine, attached internally. When the fluorescein is excited by irradiation, its fluorescent emission will be quenched if the rhodamine is close enough to be excited through the process of fluorescence energy transfer (FET).<sup>(4,5)</sup> During PCR, if the probe is hybridized to a template strand, *Taq* DNA polymerase will cleave the probe because of its inherent 5'→3' nucleolytic activity. If the cleavage occurs between the fluorescein and rhodamine dyes, it causes an increase in fluorescein fluorescence intensity because the fluorescein is no longer quenched. The increase in fluorescein fluorescence intensity indicates that the probe-specific PCR product has been generated. Thus, FET between a reporter dye and a quencher dye is critical to the performance of the probe in the 5' nuclease PCR assay.

Quenching is completely dependent on the physical proximity of the two dyes.<sup>(6)</sup> Because of this, it has been assumed that the quencher dye must be attached near the 5' end. Surprisingly, we have found that attaching a rhodamine dye at the 3' end of a probe still provides adequate quenching for the probe to perform in the 5' nuclease

PCR assay. Furthermore, cleavage of this type of probe is not required to achieve some reduction in quenching. Oligonucleotides with a reporter dye on the 5' end and a quencher dye on the 3' end exhibit a much higher reporter fluorescence when double-stranded as compared with single-stranded. This should make it possible to use this type of double-labeled probe for homogeneous detection of nucleic acid hybridization.

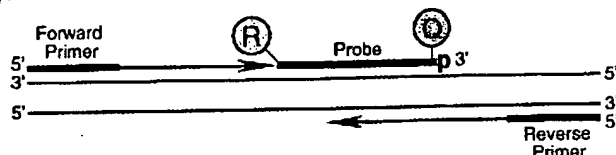
## MATERIALS AND METHODS

### Oligonucleotides

Table 1 shows the nucleotide sequence of the oligonucleotides used in this study. Linker arm nucleotide (LAN) phosphoramidite was obtained from Glen Research. The standard DNA phosphoramidites, 6-carboxyfluorescein (6-FAM) phosphoramidite, 6-carboxytetramethylrhodamine succinimidyl ester (TAMRA NHS ester), and Phosphalink for attaching a 3'-blocking phosphate, were obtained from Perkin-Elmer, Applied Biosystems Division. Oligonucleotide synthesis was performed using an ABI model 394 DNA synthesizer (Applied Biosystems). Primer and complement oligonucleotides were purified using Oligo Purification Cartridges (Applied Biosystems). Double-labeled probes were synthesized with 6-FAM-labeled phosphoramidite at the 5' end, LAN replacing one of the T's in the sequence, and Phosphalink at the 3' end. Following deprotection and ethanol precipitation, TAMRA NHS ester was coupled to the LAN-containing oligonucleotide in 250

# Research

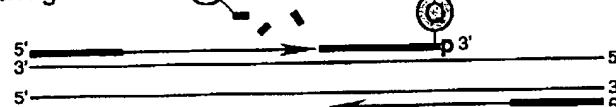
## Polymerization



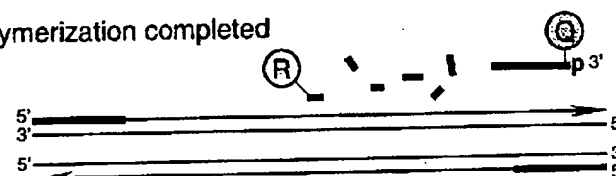
## Strand displacement



## Cleavage



## Polymerization completed



**FIGURE 1** Diagram of 5' nuclease assay. Stepwise representation of the 5' → 3' nucleolytic activity of *Taq* DNA polymerase acting on a fluorogenic probe during one extension phase of PCR.

mm Na-bicarbonate buffer (pH 9.0) at room temperature. Unreacted dye was removed by passage over a PD-10 Sephadex column. Finally, the double-labeled probe was purified by preparative high-performance liquid chromatography (HPLC) using an Aquapore C<sub>8</sub> 220×4.6-mm column with 7-μm particle size. The column was developed with a 24-min linear gradient of 8–20% acetonitrile in 0.1 M TEAA (triethylamine acetate). Probes are named by designating the sequence from Table 1 and the position of the LAN-TAMRA moiety. For example, probe A1-7 has sequence A1 with LAN-TAMRA at nucleotide position 7 from the 5' end.

### PCR Systems

All PCR amplifications were performed in the Perkin-Elmer GeneAmp PCR System 9600 using 50-μl reactions that contained 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 200 μM dATP, 200 μM dCTP, 200 μM dGTP, 400 μM dUTP, 0.5 unit of AmpErase uracil N-glycosylase (Perkin-Elmer), and 1.25 unit of AmpliTaq DNA polymerase (Perkin-Elmer). A 295-bp segment from exon 3 of the human β-actin

gene (nucleotides 2141–2435 in the sequence of Nakajima-Iijima et al.)<sup>(7)</sup> was amplified using primers AFP and ARP (Table 1), which are modified slightly from those of du Breuil et al.<sup>(8)</sup> Actin amplification reactions contained 4 mM MgCl<sub>2</sub>, 20 ng of human genomic DNA, 50 nM A1 or A3 probe, and 300 nM each

primer. The thermal regimen was 50°C (2 min), 95°C (10 min), 40 cycles of 95°C (20 sec), 60°C (1 min), and hold at 72°C. A 515-bp segment was amplified from a plasmid that consists of a segment of λ DNA (nucleotides 32,220–32,747) inserted in the *Sma*I site of vector pUC119. These reactions contained 3.5 mM MgCl<sub>2</sub>, 1 ng of plasmid DNA, 50 nM P2 or P5 probe, 200 nM primer F119, and 200 nM primer R119. The thermal regimen was 50°C (2 min), 95°C (10 min), 25 cycles of 95°C (20 sec), 57°C (1 min), and hold at 72°C.

### Fluorescence Detection

For each amplification reaction, a 40-μl aliquot of a sample was transferred to an individual well of a white, 96-well microtiter plate (Perkin-Elmer). Fluorescence was measured on the Perkin-Elmer TaqMan LS-50B System, which consists of a luminescence spectrometer with plate reader assembly, a 485-nm excitation filter, and a 515-nm emission filter. Excitation was at 488 nm using a 5-nm slit width. Emission was measured at 518 nm for 6-FAM (the reporter or R value) and 582 nm for TAMRA (the quencher or Q value) using a 10-nm slit width. To determine the increase in reporter emission that is caused by cleavage of the probe during PCR, three normalizations are applied to the raw emission data. First, emission intensity of a buffer blank is subtracted for each wavelength. Second, emission intensity of the reporter is

**TABLE 1** Sequences of Oligonucleotides

Name	Type	Sequence
F119	primer	ACCCACAGGAAGTATGATCACCAGCTC
R119	primer	ATGTCGCGTTCCGGCTGACGTTCTGC
P2	probe	TCCGATTACTGATCGTTCGCCAACCAGTp
P2C	complement	GTACTGGTTGGCAACGATCAGTAATGCGATG
P5	probe	CGGAATTGCTGGTATCTATGACAAGGATp
P5C	complement	TTGATCCTTGTGTCATAGATACCAAGCAATCCG
AFP	primer	TCACCCACACTGTGCCCCTCTACGA
ARP	primer	CAGCGGAACCGCTCATTGCCAATGG
A1	probe	ATGCCCTCCCCCATGCCATCCTGCGTp
A1C	complement	AGACGCAGGATGGCATGGGGGAGGGGCATAC
A3	probe	CGCCCTGGACTTCGAGCAAGAGATp
A3C	complement	CCATCTCTTGCTCGAAGTCCAGGGCGAC

For each oligonucleotide used in this study, the nucleic acid sequence is given, written in the 5' → 3' direction. There are three types of oligonucleotides: PCR primer, fluorogenic probe used in the 5' nuclease assay, and complement used to hybridize to the corresponding probe. For the probes, the underlined base indicates a position where LAN with TAMRA attached was substituted for a T. (p) The presence of a 3' phosphate on each probe.

A1-2 RAQGCCCTCCCCCATGCCATCCTGCGTp  
 A1-7 RATGCCCTCCCCCATGCCATCCTGCGTp  
 A1-14 RATGCCCTCCCCCAQGCCATCCTGCGTp  
 A1-19 RATGCCCTCCCCCATGCCAQCTGCGTp  
 A1-22 RATGCCCTCCCCCATGCCATCCQCGTp  
 A1-26 RATGCCCTCCCCCATGCCATCCTGCGQp

Probe	518 nm		582 nm		RQ <sup>-</sup>	RQ <sup>+</sup>	ΔRQ
	no temp.	+ temp.	no temp.	+ temp.			
A1-2	25.5 ± 2.1	32.7 ± 1.9	38.2 ± 3.0	38.2 ± 2.0	0.67 ± 0.01	0.86 ± 0.06	0.19 ± 0.06
A1-7	53.5 ± 4.3	395.1 ± 21.4	108.5 ± 6.3	110.3 ± 5.3	0.49 ± 0.03	3.58 ± 0.17	3.08 ± 0.18
A1-14	127.0 ± 4.9	403.5 ± 19.1	108.7 ± 5.3	93.1 ± 6.3	1.16 ± 0.02	4.34 ± 0.15	3.18 ± 0.15
A1-19	187.5 ± 17.9	422.7 ± 7.7	70.3 ± 7.4	73.0 ± 2.8	2.67 ± 0.05	5.80 ± 0.15	3.13 ± 0.16
A1-22	224.6 ± 9.4	482.2 ± 43.6	100.0 ± 4.0	96.2 ± 9.6	2.25 ± 0.03	5.02 ± 0.11	2.77 ± 0.12
A1-26	160.2 ± 8.9	454.1 ± 18.4	93.1 ± 5.4	90.7 ± 3.2	1.72 ± 0.02	5.01 ± 0.08	3.29 ± 0.08

**FIGURE 2** Results of 5' nuclease assay comparing β-actin probes with TAMRA at different nucleotide positions. As described in Materials and Methods, PCR amplifications containing the indicated probes were performed, and the fluorescence emission was measured at 518 and 582 nm. Reported values are the average ± 1 S.D. for six reactions run without added template (no temp.) and six reactions run with template (+ temp.). The RQ ratio was calculated for each individual reaction and averaged to give the reported RQ<sup>-</sup> and RQ<sup>+</sup> values.

divided by the emission intensity of the quencher to give an RQ ratio for each reaction tube. This normalizes for well-to-well variations in probe concentration and fluorescence measurement. Finally, ΔRQ is calculated by subtracting the RQ value of the no-template control (RQ<sup>-</sup>) from the RQ value for the complete reaction including template (RQ<sup>+</sup>).

## RESULTS

A series of probes with increasing distances between the fluorescein reporter and rhodamine quencher were tested to investigate the minimum and maximum spacing that would give an acceptable performance in the 5' nuclease PCR assay. These probes hybridize to a target

sequence in the human β-actin gene. Figure 2 shows the results of an experiment in which these probes were included in PCR that amplified a segment of the β-actin gene containing the target sequence. Performance in the 5' nuclease PCR assay is monitored by the magnitude of ΔRQ, which is a measure of the increase in reporter fluorescence caused by PCR amplification of the probe target. Probe A1-2 has a ΔRQ value that is close to zero, indicating that the probe was not cleaved appreciably during the amplification reaction. This suggests that with the quencher dye on the second nucleotide from the 5' end, there is insufficient room for *Taq* polymerase to cleave efficiently between the reporter and quencher. The other five probes exhibited comparable ΔRQ values that are

clearly different from zero. Thus, all five probes are being cleaved during PCR amplification resulting in a similar increase in reporter fluorescence. It should be noted that complete digestion of a probe produces a much larger increase in reporter fluorescence than that observed in Figure 2 (data not shown). Thus, even in reactions where amplification occurs, the majority of probe molecules remain uncleaved. It is mainly for this reason that the fluorescence intensity of the quencher dye TAMRA changes little with amplification of the target. This is what allows us to use the 582-nm fluorescence reading as a normalization factor.

The magnitude of RQ<sup>-</sup> depends mainly on the quenching efficiency inherent in the specific structure of the probe and the purity of the oligonucleotide. Thus, the larger RQ<sup>-</sup> values indicate that probes A1-14, A1-19, A1-22, and A1-26 probably have reduced quenching as compared with A1-7. Still, the degree of quenching is sufficient to detect a highly significant increase in reporter fluorescence when each of these probes is cleaved during PCR.

To further investigate the ability of TAMRA on the 3' end to quench 6-FAM on the 5' end, three additional pairs of probes were tested in the 5' nuclease PCR assay. For each pair, one probe has TAMRA attached to an internal nucleotide and the other has TAMRA attached to the 3' end nucleotide. The results are shown in Table 2. For all three sets, the probe with the 3' quencher exhibits a ΔRQ value that is considerably higher than for the probe with the internal quencher. The RQ<sup>-</sup> values suggest that differences in quenching are not as great as those observed with some of the A1 probes. These results demonstrate that a quencher dye on the 3' end of an oligonucleotide can quench efficiently the

**TABLE 2** Results of 5' Nuclease Assay Comparing Probes with TAMRA Attached to an Internal or 3'-terminal Nucleotide

Probe	518 nm		582 nm		RQ <sup>-</sup>	RQ <sup>+</sup>	ΔRQ
	no temp.	+ temp.	no temp.	+ temp.			
A3-6	54.6 ± 3.2	84.8 ± 3.7	116.2 ± 6.4	115.6 ± 2.5	0.47 ± 0.02	0.73 ± 0.03	0.26 ± 0.04
A3-24	72.1 ± 2.9	236.5 ± 11.1	84.2 ± 4.0	90.2 ± 3.8	0.86 ± 0.02	2.62 ± 0.05	1.76 ± 0.05
P2-7	82.8 ± 4.4	384.0 ± 34.1	105.1 ± 6.4	120.4 ± 10.2	0.79 ± 0.02	3.19 ± 0.16	2.40 ± 0.16
P2-27	113.4 ± 6.6	555.4 ± 14.1	140.7 ± 8.5	118.7 ± 4.8	0.81 ± 0.01	4.68 ± 0.10	3.88 ± 0.10
P5-10	77.5 ± 6.5	244.4 ± 15.9	86.7 ± 4.3	95.8 ± 6.7	0.89 ± 0.05	2.55 ± 0.06	1.66 ± 0.08
P5-28	64.0 ± 5.2	333.6 ± 12.1	100.6 ± 6.1	94.7 ± 6.3	0.63 ± 0.02	3.53 ± 0.12	2.89 ± 0.13

Reactions containing the indicated probes and calculations were performed as described in Material and Methods and in the legend to Fig. 2.

## Research

fluorescence of a reporter dye on the 5' end. The degree of quenching is sufficient for this type of oligonucleotide to be used as a probe in the 5' nuclease PCR assay.

To test the hypothesis that quenching by a 3' TAMRA depends on the flexibility of the oligonucleotide, fluorescence was measured for probes in the single-stranded and double-stranded states. Table 3 reports the fluorescence observed at 518 and 582 nm. The relative degree of quenching is assessed by calculating the RQ ratio. For probes with TAMRA 6–10 nucleotides from the 5' end, there is little difference in the RQ values when comparing single-stranded with double-stranded oligonucleotides. The results for probes with TAMRA at the 3' end are much different. For these probes, hybridization to a complementary strand causes a dramatic increase in RQ. We propose that this loss of quenching is caused by the rigid structure of double-stranded DNA, which prevents the 5' and 3' ends from being in proximity.

When TAMRA is placed toward the 3' end, there is a marked  $Mg^{2+}$  effect on quenching. Figure 3 shows a plot of observed RQ values for the A1 series of probes as a function of  $Mg^{2+}$  concentration. With TAMRA attached near the 5' end (probe A1-2 or A1-7), the RQ value at 0 mM  $Mg^{2+}$  is only slightly higher than RQ at 10 mM  $Mg^{2+}$ . For probes A1-19, A1-22, and A1-26, the RQ values at 0 mM  $Mg^{2+}$  are very high, indicating a much

reduced quenching efficiency. For each of these probes, there is a marked decrease in RQ at 1 mM  $Mg^{2+}$  followed by a gradual decline as the  $Mg^{2+}$  concentration increases to 10 mM. Probe A1-14 shows an intermediate RQ value at 0 mM  $Mg^{2+}$  with a gradual decline at higher  $Mg^{2+}$  concentrations. In a low-salt environment with no  $Mg^{2+}$  present, a single-stranded oligonucleotide would be expected to adopt an extended conformation because of electrostatic repulsion. The binding of  $Mg^{2+}$  ions acts to shield the negative charge of the phosphate backbone so that the oligonucleotide can adopt conformations where the 3' end is close to the 5' end. Therefore, the observed  $Mg^{2+}$  effects support the notion that quenching of a 5' reporter dye by TAMRA at or near the 3' end depends on the flexibility of the oligonucleotide.

### DISCUSSION

The striking finding of this study is that it seems the rhodamine dye TAMRA, placed at any position in an oligonucleotide, can quench the fluorescent emission of a fluorescein (6-FAM) placed at the 5' end. This implies that a single-stranded, double-labeled oligonucleotide must be able to adopt conformations where the TAMRA is close to the 5' end. It should be noted that the decay of 6-FAM in the excited state requires a certain amount of time. Therefore, what

matters for quenching is not the average distance between 6-FAM and TAMRA but, rather, how close TAMRA can get to 6-FAM during the lifetime of the 6-FAM excited state. As long as the decay time of the excited state is relatively long compared with the molecular motions of the oligonucleotide, quenching can occur. Thus, we propose that TAMRA at the 3' end, or any other position, can quench 6-FAM at the 5' end because TAMRA is in proximity to 6-FAM often enough to be able to accept energy transfer from an excited 6-FAM.

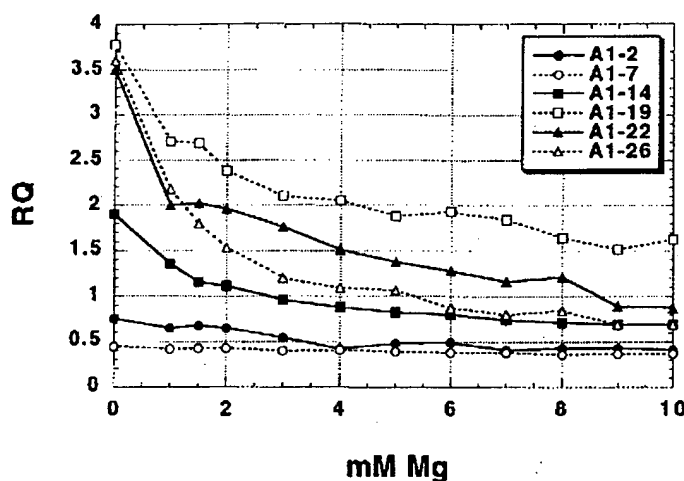
Details of the fluorescence measurements remain puzzling. For example, Table 3 shows that hybridization of probes A1-26, A3-24, and P5-28 to their complementary strands not only causes a large increase in 6-FAM fluorescence at 518 nm but also causes a modest increase in TAMRA fluorescence at 582 nm. If TAMRA is being excited by energy transfer from quenched 6-FAM, then loss of quenching attributable to hybridization should cause a decrease in the fluorescence emission of TAMRA. The fact that the fluorescence emission of TAMRA increases indicates that the situation is more complex. For example, we have anecdotal evidence that the bases of the oligonucleotide, especially G, quench the fluorescence of both 6-FAM and TAMRA to some degree. When double-stranded, base-pairing may reduce the ability of the bases to quench. The primary factor causing the quenching of 6-FAM in an intact probe is the TAMRA dye. Evidence for the importance of TAMRA is that 6-FAM fluorescence remains relatively unchanged when probes labeled only with 6-FAM are used in the 5' nuclease PCR assay (data not shown). Secondary effectors of fluorescence, both before and after cleavage of the probe, need to be explored further.

Regardless of the physical mechanism, the relative independence of position and quenching greatly simplifies the design of probes for the 5' nuclease PCR assay. There are three main factors that determine the performance of a double-labeled fluorescent probe in the 5' nuclease PCR assay. The first factor is the degree of quenching observed in the intact probe. This is characterized by the value of  $RQ^{-}$ , which is the ratio of reporter to quencher fluorescent emissions for a no template control PCR. Influences on the value of  $RQ^{-}$  include the particular reporter and quencher

**TABLE 3** Comparison of Fluorescence Emissions of Single-stranded and Double-stranded Fluorogenic Probes

Probe	518 nm		582 nm		RQ	
	ss	ds	ss	ds	ss	ds
A1-7	27.75	68.53	61.08	138.18	0.45	0.50
A1-26	43.31	509.38	53.50	93.86	0.81	5.43
A3-6	16.75	62.88	39.33	165.57	0.43	0.38
A3-24	30.05	578.64	67.72	140.25	0.45	3.21
P2-7	35.02	70.13	54.63	121.09	0.64	0.58
P2-27	39.89	320.47	65.10	61.13	0.61	5.25
P5-10	27.34	144.85	61.95	165.54	0.44	0.87
P5-28	33.65	462.29	72.39	104.61	0.46	4.43

(ss) Single-stranded. The fluorescence emissions at 518 or 582 nm for solutions containing a final concentration of 50 nM indicated probe, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, and 10 mM  $MgCl_2$ . (ds) Double-stranded. The solutions contained, in addition, 100 nM A1C for probes A1-7 and A1-26, 100 nM A3C for probes A3-6 and A3-24, 100 nM P2C for probes P2-7 and P2-27, or 100 nM P5C for probes P5-10 and P5-28. Before the addition of  $MgCl_2$ , 120  $\mu$ l of each sample was heated at 95°C for 5 min. Following the addition of 80  $\mu$ l of 25 mM  $MgCl_2$ , each sample was allowed to cool to room temperature and the fluorescence emissions were measured. Reported values are the average of three determinations.



**FIGURE 3** Effect of  $Mg^{2+}$  concentration on RQ ratio for the A1 series of probes. The fluorescence emission intensity at 518 and 582 nm was measured for solutions containing 50 nM probe, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, and varying amounts (0–10 mM) of  $MgCl_2$ . The calculated RQ ratios (518 nm intensity divided by 582 nm intensity) are plotted vs.  $MgCl_2$  concentration (mM Mg). The key (upper right) shows the probes examined.

dyes used, spacing between reporter and quencher dyes, nucleotide sequence context effects, presence of structure or other factors that reduce flexibility of the oligonucleotide, and purity of the probe. The second factor is the efficiency of hybridization, which depends on probe  $T_m$ , presence of secondary structure in probe or template, annealing temperature, and other reaction conditions. The third factor is the efficiency at which *Taq* DNA polymerase cleaves the bound probe between the reporter and quencher dyes. This cleavage is dependent on sequence complementarity between probe and template as shown by the observation that mismatches in the segment between reporter and quencher dyes drastically reduce the cleavage of probe.<sup>(1)</sup>

The rise in  $RQ^-$  values for the A1 series of probes seems to indicate that the degree of quenching is reduced somewhat as the quencher is placed toward the 3' end. The lowest apparent quenching is observed for probe A1-19 (see Fig. 3) rather than for the probe where the TAMRA is at the 3' end (A1-26). This is understandable, as the conformation of the 3' end position would be expected to be less restricted than the conformation of an internal position. In effect, a quencher at the 3' end is freer to adopt conformations close to the 5' reporter dye than is an internally placed quencher. For the other three sets of

probes, the interpretation of  $RQ^-$  values is less clear-cut. The A3 probes show the same trend as A1, with the 3' TAMRA probe having a larger  $RQ^-$  than the internal TAMRA probe. For the P2 pair, both probes have about the same  $RQ^-$  value. For the P5 probes, the  $RQ^-$  for the 3' probe is less than for the internally labeled probe. Another factor that may explain some of the observed variation is that purity affects the  $RQ^-$  value. Although all probes are HPLC purified, a small amount of contamination with unquenched reporter can have a large effect on  $RQ^-$ .

Although there may be a modest effect on degree of quenching, the position of the quencher apparently can have a large effect on the efficiency of probe cleavage. The most drastic effect is observed with probe A1-2, where placement of the TAMRA on the second nucleotide reduces the efficiency of cleavage to almost zero. For the A3, P2, and P5 probes,  $\Delta RQ$  is much greater for the 3' TAMRA probes as compared with the internal TAMRA probes. This is explained most easily by assuming that probes with TAMRA at the 3' end are more likely to be cleaved between reporter and quencher than are probes with TAMRA attached internally. For the A1 probes, the cleavage efficiency of probe A1-7 must already be quite high, as  $\Delta RQ$  does not increase when the quencher is placed closer to the 3' end. This illus-

trates the importance of being able to use probes with a quencher on the 3' end in the 5' nuclease PCR assay. In this assay, an increase in the intensity of reporter fluorescence is observed only when the probe is cleaved between the reporter and quencher dyes. By placing the reporter and quencher dyes on the opposite ends of an oligonucleotide probe, any cleavage that occurs will be detected. When the quencher is attached to an internal nucleotide, sometimes the probe works well (A1-7) and other times not so well (A3-6). The relatively poor performance of probe A3-6 presumably means the probe is being cleaved 3' to the quencher rather than between the reporter and quencher. Therefore, the best chance of having a probe that reliably detects accumulation of PCR product in the 5' nuclease PCR assay is to use a probe with the reporter and quencher dyes on opposite ends.

Placing the quencher dye on the 3' end may also provide a slight benefit in terms of hybridization efficiency. The presence of a quencher attached to an internal nucleotide might be expected to disrupt base-pairing and reduce the  $T_m$  of a probe. In fact, a 2°C–3°C reduction in  $T_m$  has been observed for two probes with internally attached TAMRAs.<sup>(9)</sup> This disruptive effect would be minimized by placing the quencher at the 3' end. Thus, probes with 3' quenchers might exhibit slightly higher hybridization efficiencies than probes with internal quenchers.

The combination of increased cleavage and hybridization efficiencies means that probes with 3' quenchers probably will be more tolerant of mismatches between probe and target as compared with internally labeled probes. This tolerance of mismatches can be advantageous, as when trying to use a single probe to detect PCR-amplified products from samples of different species. Also, it means that cleavage of probe during PCR is less sensitive to alterations in annealing temperature or other reaction conditions. The one application where tolerance of mismatches may be a disadvantage is for allelic discrimination. Lee et al.<sup>(1)</sup> demonstrated that allele-specific probes were cleaved between reporter and quencher only when hybridized to a perfectly complementary target. This allowed them to distinguish the normal human cystic fibrosis allele from the  $\Delta F508$  mutant. Their probes had TAMRA attached to the seventh nucleotide from

## Research

the 5' end and were designed so that any mismatches were between the reporter and quencher. Increasing the distance between reporter and quencher would lessen the disruptive effect of mismatches and allow cleavage of the probe on the incorrect target. Thus, probes with a quencher attached to an internal nucleotide may still be useful for allelic discrimination.

In this study loss of quenching upon hybridization was used to show that quenching by a 3' TAMRA is dependent on the flexibility of a single-stranded oligonucleotide. The increase in reporter fluorescence intensity, though, could also be used to determine whether hybridization has occurred or not. Thus, oligonucleotides with reporter and quencher dyes attached at opposite ends should also be useful as hybridization probes. The ability to detect hybridization in real time means that these probes could be used to measure hybridization kinetics. Also, this type of probe could be used to develop homogeneous hybridization assays for diagnostics or other applications. Bagwell et al.<sup>(10)</sup> describe just this type of homogeneous assay where hybridization of a probe causes an increase in fluorescence caused by a loss of quenching. However, they utilized a complex probe design that requires adding nucleotides to both ends of the probe sequence to form two imperfect hairpins. The results presented here demonstrate that the simple addition of a reporter dye to one end of an oligonucleotide and a quencher dye to the other end generates a fluorogenic probe that can detect hybridization or PCR amplification.

### ACKNOWLEDGMENTS

We acknowledge Lincoln McBride of Perkin-Elmer for his support and encouragement on this project and Mitch Winnik of the University of Toronto for helpful discussions on time-resolved fluorescence.

### REFERENCES

1. Lee, L.G., C.R. Connell, and W. Bloch. 1993. Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Res.* **21**: 3761-3766.
2. Holland, P.M., R.D. Abramson, R. Watson, and D.H. Gelfand. 1991. Detection of specific polymerase chain reaction product by utilizing the 5' to 3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. Natl. Acad. Sci.* **88**: 7276-7280.
3. Lyamichev, V., M.A.D. Brow, and J.E. Dahlberg. 1993. Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science* **260**: 778-783.
4. Förster, V.Th. 1948. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys. (Leipzig)* **2**: 55-75.
5. Lakowicz, J.R. 1983. Energy transfer. In *Principles of fluorescent spectroscopy*, pp. 303-339. Plenum Press, New York, NY.
6. Stryer, L. and R.P. Haugland. 1967. Energy transfer: A spectroscopic ruler. *Proc. Natl. Acad. Sci.* **58**: 719-726.
7. Nakajima-Iijima, S., H. Hamada, P. Reddy, and T. Kakunaga. 1985. Molecular structure of the human cytoplasmic beta-actin gene: Inter-species homology of sequences in the introns. *Proc. Natl. Acad. Sci.* **82**: 6133-6137.
8. du Breuil, R.M., J.M. Patel, and B.V. Mendelow. 1993. Quantitation of  $\beta$ -actin-specific mRNA transcripts using xeno-competitive PCR. *PCR Methods Applic.* **3**: 57-59.
9. Livak, K.J. (unpubl.).
10. Bagwell, C.B., M.E. Munson, R.L. Christensen, and E.J. Lovett. 1994. A new homogeneous assay system for specific nucleic acid sequences: Poly-dA and poly-A detection. *Nucleic Acids Res.* **22**: 2424-2425.

Received December 20, 1994; accepted in revised form March 6, 1995.

THIS MATERIAL MAY BE PROTECTED  
BY COPYRIGHT LAW (17 U.S. CODE)

## GENOME METHODS

## Real Time Quantitative PCR

Christian A. Heid,<sup>1</sup> Junko Stevens,<sup>2</sup> Kenneth J. Livak,<sup>2</sup> and  
P. Mickey Williams<sup>1,3</sup>

<sup>1</sup>BioAnalytical Technology Department, Genentech, Inc., South San Francisco, California 94080;

<sup>2</sup>Applied BioSystems Division of Perkin Elmer Corp., Foster City, California 94404

We have developed a novel "real time" quantitative PCR method. The method measures PCR product accumulation through a dual-labeled fluorogenic probe (i.e., TaqMan Probe). This method provides very accurate and reproducible quantitation of gene copies. Unlike other quantitative PCR methods, real-time PCR does not require post-PCR sample handling, preventing potential PCR product carry-over contamination and resulting in much faster and higher throughput assays. The real-time PCR method has a very large dynamic range of starting target molecule determination (at least five orders of magnitude). Real-time quantitative PCR is extremely accurate and less labor-intensive than current quantitative PCR methods.

Quantitative nucleic acid sequence analysis has had an important role in many fields of biological research. Measurement of gene expression (RNA) has been used extensively in monitoring biological responses to various stimuli (Tan et al. 1994; Huang et al. 1995a,b; Prud'homme et al. 1995). Quantitative gene analysis (DNA) has been used to determine the genomic quantity of a particular gene, as in the case of the human *HER2* gene, which is amplified in ~30% of breast tumors (Slamon et al. 1987). Gene and genome quantitation (DNA and RNA) also have been used for analysis of human immunodeficiency virus (HIV) burden demonstrating changes in the levels of virus throughout the different phases of the disease (Connor et al. 1993; Platak et al. 1993b; Furtado et al. 1995).

Many methods have been described for the quantitative analysis of nucleic acid sequences (both for RNA and DNA; Southern 1975; Sharp et al. 1980; Thomas 1980). Recently, PCR has proven to be a powerful tool for quantitative nucleic acid analysis. PCR and reverse transcriptase (RT)-PCR have permitted the analysis of minimal starting quantities of nucleic acid (as little as one cell equivalent). This has made possible many experiments that could not have been performed with traditional methods. Although PCR has provided a powerful tool, it is imperative

that it be used properly for quantitation (Rasmussen 1995). Many early reports of quantitative PCR and RT-PCR described quantitation of the PCR product but did not measure the initial target sequence quantity. It is essential to design proper controls for the quantitation of the initial target sequences (Perre 1992; Clementi et al. 1993).

Researchers have developed several methods of quantitative PCR and RT-PCR. One approach measures PCR product quantity in the log phase of the reaction before the plateau (Kellogg et al. 1990; Pang et al. 1990). This method requires that each sample has equal input amounts of nucleic acid and that each sample under analysis amplifies with identical efficiency up to the point of quantitative analysis. A gene sequence (contained in all samples at relatively constant quantities, such as  $\beta$ -actin) can be used for sample amplification efficiency normalization. Using conventional methods of PCR detection and quantitation (gel electrophoresis or plate capture hybridization), it is extremely laborious to assure that all samples are analyzed during the log phase of the reaction (for both the target gene and the normalization gene). Another method, quantitative competitive (QC)-PCR, has been developed and is used widely for PCR quantitation. QC-PCR relies on the inclusion of an internal control competitor in each reaction (Becker-Andre 1991; Platak et al. 1993a,b). The efficiency of each reaction is normalized to the internal competitor. A known amount of internal competitor can be

<sup>3</sup>Corresponding author.

## REAL TIME QUANTITATIVE PCR

## RESULTS

## PCR Product Detection in Real Time

The goal was to develop a high-throughput, sensitive, and accurate gene quantitation assay for use in monitoring lipid mediated therapeutic gene delivery. A plasmid encoding human factor VIII gene sequence, pF8TM (see Methods), was used as a model therapeutic gene. The assay uses fluorescent Taqman methodology and an instrument capable of measuring fluorescence in real time (ABI Prism 7700 Sequence Detector). The Taqman reaction requires a hybridization probe labeled with two different fluorescent dyes. One dye is a reporter dye (FAM), the other is a quenching dye (TAMRA). When the probe is intact, fluorescent energy transfer occurs and the reporter dye fluorescent emission is absorbed by the quenching dye (TAMRA). During the extension phase of the PCR cycle, the fluorescent hybridization probe is cleaved by the 5'-3' nucleolytic activity of the DNA polymerase. On cleavage of the probe, the reporter dye emission is no longer transferred efficiently to the quenching dye, resulting in an increase of the reporter dye fluorescent emission spectra. PCR primers and probes were designed for the human factor VIII sequence and human  $\beta$ -actin gene (as described in Methods). Optimization reactions were performed to choose the appropriate probe and magnesium concentrations yielding the highest intensity of reporter fluorescent signal without sacrificing specificity. The instrument uses a charge-coupled device (i.e., CCD camera) for measuring the fluorescent emission spectra from 500 to 650 nm. Each PCR tube was monitored sequentially for 25 msec with continuous monitoring throughout the amplification. Each tube was re-examined every 8.5 sec. Computer software was designed to examine the fluorescent intensity of both the reporter dye (FAM) and the quenching dye (TAMRA). The fluorescent intensity of the quenching dye, TAMRA, changes very little over the course of the PCR amplification (data not shown). Therefore, the intensity of TAMRA dye emission serves as an internal standard with which to normalize the reporter dye (FAM) emission variations. The software calculates a value termed  $\Delta Rn$  (or  $\Delta RQ$ ) using the following equation:  $\Delta Rn = (Rn^t - Rn^i) / (Rn^t)$ , where  $Rn^t$  = emission intensity of reporter/emission intensity of quencher at any given time in a reaction tube, and  $Rn^i$  = emission intensity of re-

added to each sample. To obtain relative quantitation, the unknown target PCR product is compared with the known competitor PCR product. Success of a quantitative competitive PCR assay relies on developing an internal control that amplifies with the same efficiency as the target molecule. The design of the competitor and the validation of amplification efficiencies require a dedicated effort. However, because QC-PCR does not require that PCR products be analyzed during the log phase of the amplification, it is the easier of the two methods to use.

Several detection systems are used for quantitative PCR and RT-PCR analysis: (1) agarose gels, (2) fluorescent labeling of PCR products and detection with laser-induced fluorescence using capillary electrophoresis (Fusco et al. 1995; Williams et al. 1996) or acrylamide gels, and (3) plate capture and sandwich probe hybridization (Mulder et al. 1994). Although these methods proved successful, each method requires post-PCR manipulations that add time to the analysis and may lead to laboratory contamination. The sample throughput of these methods is limited (with the exception of the plate capture approach), and, therefore, these methods are not well suited for uses demanding high sample throughput (i.e., screening of large numbers of biomolecules or analyzing samples for diagnostics or clinical trials).

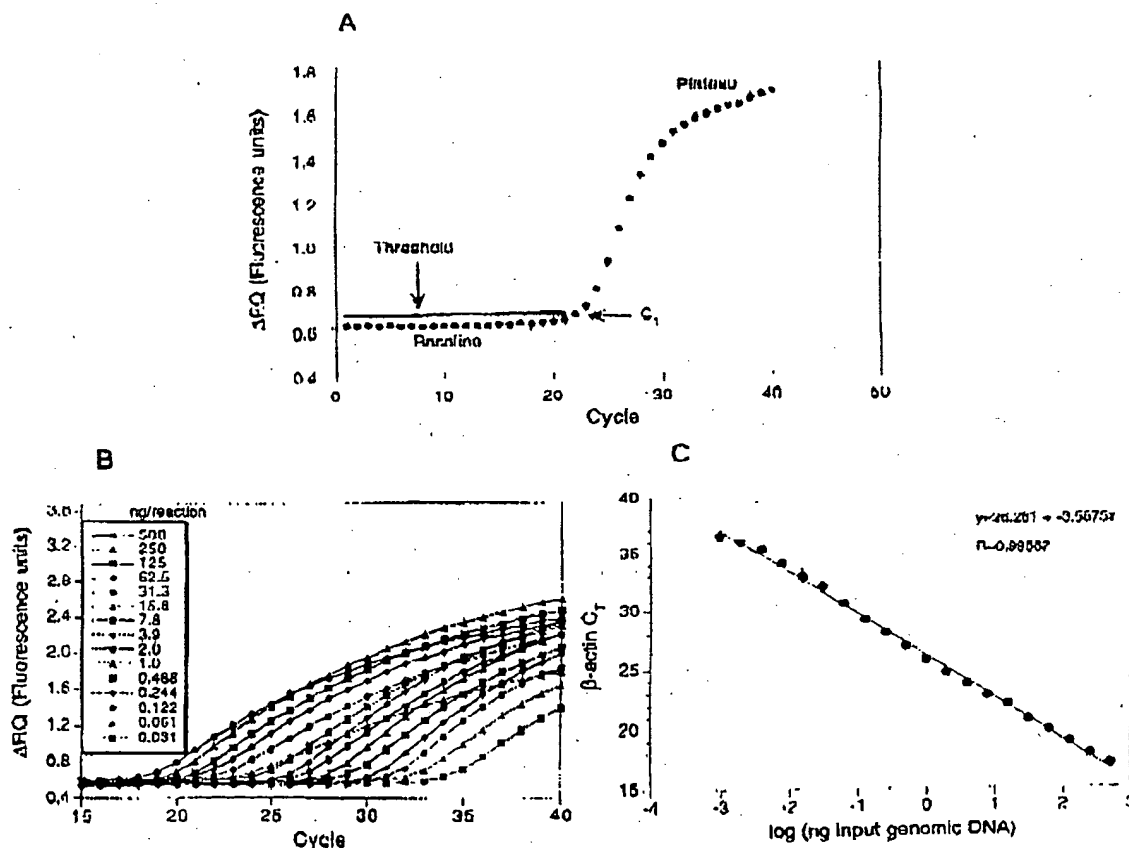
Here we report the development of a novel assay for quantitative DNA analysis. The assay is based on the use of the 5' nuclease assay first described by Holland et al. (1991). The method uses the 5' nuclease activity of *Taq* polymerase to cleave a nonextendible hybridization probe during the extension phase of PCR. The approach uses dual-labeled fluorogenic hybridization probes (Lee et al. 1993; Bassler et al. 1995; Livak et al. 1995a,b). One fluorescent dye serves as a reporter [FAM (i.e., 6-carboxyfluorescein)] and its emission spectra is quenched by the second fluorescent dye, TAMRA (i.e., 6-carboxy-tetramethylrhodamine). The nuclease degradation of the hybridization probe releases the quenching of the FAM fluorescent emission, resulting in an increase in peak fluorescent emission at 518 nm. The use of a sequence detector (ABI Prism) allows measurement of fluorescent spectra of all 96 wells of the thermal cycler continuously during the PCR amplification. Therefore, the reactions are monitored in real time. The output data is described and quantitative analysis of input target DNA sequences is discussed below.



## HUI ET AL.

porter/emission intensity of quencher measured prior to PCR amplification in that same reaction tube. For the purpose of quantitation, the last three data points ( $\Delta Rn$ s) collected during the extension step for each PCR cycle were analyzed. The nucleolytic degradation of the hybridization probe occurs during the extension phase of PCR, and, therefore, reporter fluorescent emission increases during this time. The three data points were averaged for each PCR cycle and the mean value for each was plotted in an "amplification plot" shown in Figure 1A. The  $\Delta Rn$  mean value is plotted on the y-axis, and time, represented by cycle number, is plotted on the x-axis. During the early cycles of the PCR amplification, the  $\Delta Rn$

value remains at base line. When sufficient hybridization probe has been cleaved by the *Taq* polymerase nuclease activity, the intensity of reporter fluorescent emission increases. Most PCR amplifications reach a plateau phase of reporter fluorescent emission if the reaction is carried out to high cycle numbers. The amplification plot is examined early in the reaction, at a point that represents the log phase of product accumulation. This is done by assigning an arbitrary threshold that is based on the variability of the base-line data. In Figure 1A, the threshold was set at 10 standard deviations above the mean of base line emission calculated from cycles 1 to 15. Once the threshold is chosen, the point at which



**Figure 1** PCR product detection in real time. (A) The Model 7700 software will construct amplification plots from the extension phase fluorescent emission data collected during the PCR amplification. The standard deviation is determined from the data points collected from the base line of the amplification plot.  $C_T$  values are calculated by determining the point at which the fluorescence exceeds a threshold limit (usually 10 times the standard deviation of the base line). (B) Overlay of amplification plots of serially (1:2) diluted human genomic DNA samples amplified with  $\beta$ -actin primers. (C) Input DNA concentration of the samples plotted versus  $C_T$ . All

## REAL TIME QUANTITATIVE PCR

the amplification plot crosses the threshold is defined as  $C_T$ .  $C_T$  is reported as the cycle number at this point. As will be demonstrated, the  $C_T$  value is predictive of the quantity of input target.

#### $C_T$ Values Provide a Quantitative Measurement of Input Target Sequences

Figure 1B shows amplification plots of 15 different PCR amplifications overlaid. The amplifications were performed on a 1:2 serial dilution of human genomic DNA. The amplified target was human  $\beta$  actin. The amplification plots shift to the right (to higher threshold cycles) as the input target quantity is reduced. This is expected because reactions with fewer starting copies of the target molecule require greater amplification to degrade enough probe to attain the threshold fluorescence. An arbitrary threshold of 10 standard deviations above the base line was used to determine the  $C_T$  values. Figure 1C represents the  $C_T$  values plotted versus the sample dilution value. Each dilution was amplified in triplicate PCR amplifications and plotted as mean values with error bars representing one standard deviation. The  $C_T$  values decrease linearly with increasing target quantity. Thus,  $C_T$  values can be used as a quantitative measurement of the input target number. It should be noted that the amplification plot for the 15.6-ng sample shown in Figure 1B does not reflect the same fluorescent rate of increase exhibited by most of the other samples. The 15.6-ng sample also achieves endpoint plateau at a lower fluorescent value than would be expected based on the input DNA. This phenomenon has been observed occasionally with other samples (data not shown) and may be attributable to late cycle inhibition; this hypothesis is still under investigation. It is important to note that the flattened slope and early plateau do not impact significantly the calculated  $C_T$  value as demonstrated by the fit on the line shown in Figure 1C. All triplicate amplifications resulted in very similar  $C_T$  values—the standard deviation did not exceed 0.5 for any dilution. This experiment contains a >100,000-fold range of input target molecules. Using  $C_T$  values for quantitation permits a much larger assay range than directly using total fluorescent emission intensity for quantitation. The linear range of fluorescent intensity measurement of the ABI Prism 7700 Se-

ments over a very large range of relative starting target quantities.

#### Sample Preparation Validation

Several parameters influence the efficiency of PCR amplification: magnesium and salt concentrations, reaction conditions (i.e., time and temperature), PCR target size and composition, primer sequences, and sample purity. All of the above factors are common to a single PCR assay, except sample to sample purity. In an effort to validate the method of sample preparation for the factor VIII assay, PCR amplification reproducibility and efficiency of 10 replicate sample preparations were examined. After genomic DNA was prepared from the 10 replicate samples, the DNA was quantitated by ultraviolet spectroscopy. Amplifications were performed analyzing  $\beta$ -actin gene content in 100 and 25 ng of total genomic DNA. Each PCR amplification was performed in triplicate. Comparison of  $C_T$  values for each triplicate sample show minimal variation based on standard deviation and coefficient of variance (Table 1). Therefore, each of the triplicate PCR amplifications was highly reproducible, demonstrating that real time PCR using this instrumentation introduces minimal variation into the quantitative PCR analysis. Comparison of the mean  $C_T$  values of the 10 replicate sample preparations also showed minimal variability, indicating that each sample preparation yielded similar results for  $\beta$ -actin gene quantity. The highest  $C_T$  difference between any of the samples was 0.85 and 0.71 for the 100 and 25 ng samples, respectively. Additionally, the amplification of each sample exhibited an equivalent rate of fluorescent emission intensity change per amount of DNA target analyzed as indicated by similar slopes derived from the sample dilutions (Fig. 2). Any sample containing an excess of a PCR inhibitor would exhibit a greater measured  $\beta$ -actin  $C_T$  value for a given quantity of DNA. In addition, the inhibitor would be diluted along with the sample in the dilution analysis (Fig. 2), altering the expected  $C_T$  value change. Each sample amplification yielded a similar result in the analysis, demonstrating that this method of sample preparation is highly reproducible with regard to sample purity.

#### Quantitative Analysis of a Plasmid After

7000 001 050 VVJ RC:BT 7007/00/71

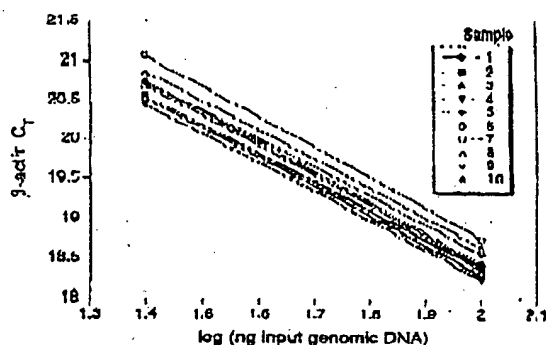
## III D F AL

Table 1. Reproducibility of Sample Preparation Method

Sample no.	100 ng				25 ng			
	C <sub>T</sub>	mean	standard deviation	CV	C <sub>T</sub>	mean	standard deviation	CV
1	18.24	18.27	0.06	0.32	20.48	20.51	0.03	0.17
	18.23				20.55			
	18.33				20.5			
2	18.33	18.37	0.06	0.32	20.61	20.54	0.11	0.54
	18.35				20.59			
	18.44				20.41			
3	18.3	18.34	0.07	0.36	20.54	20.54	0.06	0.28
	18.3				20.6			
	18.42				20.49			
4	18.15	18.23	0.08	0.46	20.48	20.43	0.05	0.26
	18.23				20.44			
	18.32				20.38			
5	18.4	18.42	0.04	0.23	20.68	20.73	0.13	0.61
	18.38				20.87			
	18.46				20.63			
6	18.54	18.74	0.24	1.26	21.09	21.06	0.03	0.15
	18.67				21.04			
	19				21.04			
7	18.28	18.39	0.12	0.66	20.67	20.68	0.04	0.2
	18.36				20.73			
	18.52				20.65			
8	18.45	18.63	0.16	0.83	20.98	20.86	0.12	0.57
	18.7				20.84			
	18.73				20.75			
9	18.18	18.29	0.1	0.55	20.46	20.51	0.07	0.32
	18.34				20.54			
	18.36				20.48			
10	18.42	18.55	0.12	0.65	20.79	20.73	0.1	0.16
	18.57				20.78			
	18.66				20.62			
Mean	(1 10)	18.42	0.17	0.90		20.66	0.19	0.94

(or containing a partial cDNA for human factor VIII, pF8TM. A series of transfections was set up using a decreasing amount of the plasmid (40, 4, 0.5, and 0.1 µg). Twenty-four hours post-transfection, total DNA was purified from each flask of cells. β-Actin gene quantity was chosen as a value for normalization of genomic DNA concentration from each sample. In this experiment, β-actin gene content should remain constant relative to total genomic DNA. Figure 3 shows the result of the β-actin DNA measurement (100 ng total DNA determined by ultraviolet spectroscopy) of each sample. Each sample was analyzed in triplicate and the mean β-actin C<sub>T</sub> values of the triplicates were plotted (error bars represent one standard deviation). The highest difference

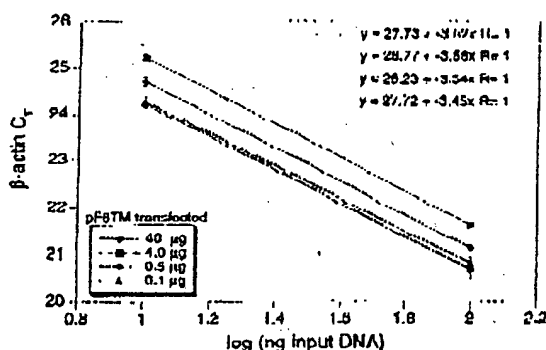
between any two sample means was 0.95 C<sub>T</sub>. Ten nanograms of total DNA of each sample were also examined for β-actin. The results again showed that very similar amounts of genomic DNA were present; the maximum mean β-actin C<sub>T</sub> value difference was 1.0. As Figure 3 shows, the rate of β-actin C<sub>T</sub> change between the 100 and 10-ng samples was similar (slope values range between 3.56 and -3.45). This verifies again that the method of sample preparation yields samples of identical PCR integrity (i.e., no sample contained an excessive amount of a PCR inhibitor). However, these results indicate that each sample contained slight differences in the actual amount of genomic DNA analyzed. Determination of actual genomic DNA concentration was accomplished



**Figure 2** Sample preparation purity. The replicate samples shown in Table 1 were also amplified in triplicate using 25 ng of each DNA sample. The figure shows the input DNA concentration (100 and 25 ng) vs.  $C_T$ . In the figure, the 100 and 25 ng points for each sample are connected by a line.

by plotting the mean  $\beta$ -actin  $C_T$  value obtained for each 100-ng sample on a  $\beta$ -actin standard curve (shown in Fig. 4C). The actual genomic DNA concentration of each sample,  $a$ , was obtained by extrapolation to the x-axis.

Figure 4A shows the measured (i.e., non-normalized) quantities of factor VIII plasmid DNA (pF8TM) from each of the four transient cell transfections. Each reaction contained 100 ng of total sample DNA (as determined by UV spectroscopy). Each sample was analyzed in triplicate



**Figure 3** Analysis of transfected cell DNA quantity and purity. The DNA preparations of the four 293 cell transfections (40, 4, 0.5, and 0.1  $\mu$ g of pF8TM) were analyzed for the  $\beta$ -actin gene. 100 and 10 ng (determined by ultraviolet spectroscopy) of each sample were amplified in triplicate. For each amount of pF8TM that was transfected, the  $\beta$ -actin  $C_T$  values are plotted versus the total input DNA concentration.

## REAL TIME QUANTITATIVE PCR

PCR amplifications. As shown, pF8TM purified from the 293 cells decreases (mean  $C_T$  values increase) with decreasing amounts of plasmid transfected. The mean  $C_T$  values obtained for pF8TM in Figure 4A were plotted on a standard curve comprised of serially diluted pF8TM, shown in Figure 4B. The quantity of pF8TM,  $b$ , found in each of the four transfections was determined by extrapolation to the x-axis of the standard curve in Figure 4B. These uncorrected values,  $b$ , for pF8TM were normalized to determine the actual amount of pF8TM found per 100 ng of genomic DNA by using the equation:

$$\frac{b \times 100 \text{ ng}}{a} = \text{actual pF8TM copies per 100 ng of genomic DNA}$$

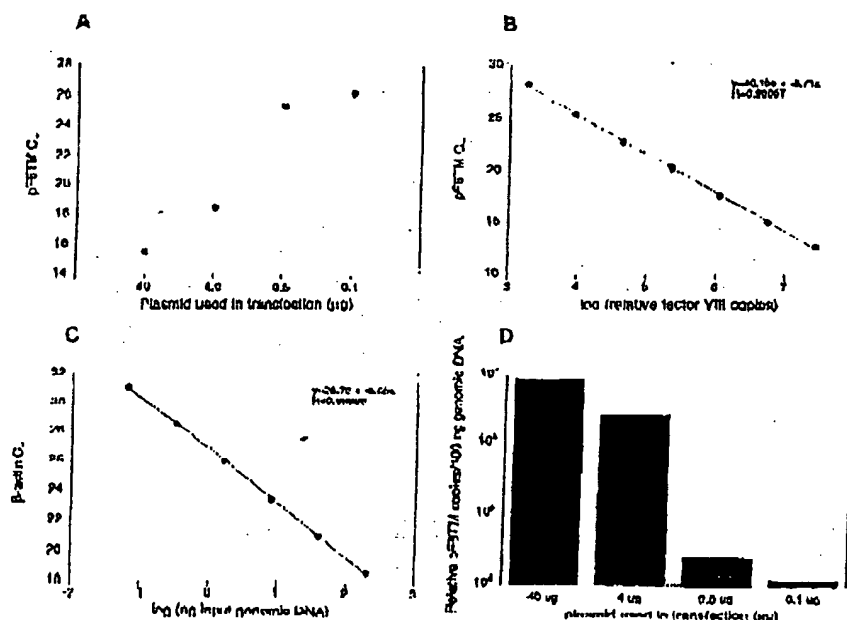
where  $a$  = actual genomic DNA in a sample and  $b$  = pF8TM copies from the standard curve. The normalized quantity of pF8TM per 100 ng of genomic DNA for each of the four transfections is shown in Figure 4D. These results show that the quantity of factor VIII plasmid associated with the 293 cells, 24 hr after transfection, decreases with decreasing plasmid concentration used in the transfection. The quantity of pF8TM associated with 293 cells, after transfection with 40  $\mu$ g of plasmid, was 35 pg per 100 ng genomic DNA. This results in ~520 plasmid copies per cell.

## DISCUSSION

We have described a new method for quantitating gene copy numbers using real-time analysis of PCR amplifications. Real-time PCR is compatible with either of the two PCR (RT-PCR) approaches: (1) quantitative competitive where an internal competitor for each target sequence is used for normalization (data not shown) or (2) quantitative comparative PCR using a normalization gene contained within the sample (i.e.,  $\beta$ -actin) or a "housekeeping" gene for RT-PCR. If equal amounts of nucleic acid are analyzed for each sample and if the amplification efficiency before quantitative analysis is identical for each sample, the internal control (normalization gene or competitor) should give equal signals for all samples.

The real-time PCR method offers several advantages over the other two methods currently employed (see the Introduction). First, the real-time PCR method is performed in a closed-tube system and requires no post-PCR manipulation

HLID L1 AL



**Figure 4** Quantitative analysis of pF8TM in transfected cells. (A) Amount of plasmid DNA used for the transfection plotted against the mean  $C_1$  value determined for pF8TM remaining 24 hr after transfection. (B,C) Standard curves of pF8TM and  $\beta$ -actin, respectively. pF8TM DNA (B) and genomic DNA (C) were diluted serially 1:5 before amplification with the appropriate primers. The  $\beta$ -actin standard curve was used to normalize the results of A to 100 ng of genomic DNA. (D) The amount of pF8TM present per 100 ng of genomic DNA.

of sample. Therefore, the potential for PCR contamination in the laboratory is reduced because amplified products can be analyzed and disposed of without opening the reaction tubes. Second, this method supports the use of a normalization gene (i.e.,  $\beta$ -actin) for quantitative PCR or house-keeping genes for quantitative RT-PCR controls. Analysis is performed in real time during the log phase of product accumulation. Analysis during log phase permits many different genes (over a wide input target range) to be analyzed simultaneously, without concern of reaching reaction plateau at different cycles. This will make multi-gene analysis assays much easier to develop, because individual internal competitors will not be needed for each gene under analysis. Third, sample throughput will increase dramatically with the new method because there is no post-PCR processing time. Additionally, working in a 96-well format is highly compatible with automation technology.

The real-time PCR method is highly reproducible. Replicate amplifications can be analyzed

for each sample minimizing potential error. The system allows for a very large assay dynamic range (approaching 1,000,000-fold starting target). Using a standard curve for the target of interest, relative copy number values can be determined for any unknown sample. Fluorescent threshold values,  $C_p$ , correlate linearly with relative DNA copy numbers. Real time quantitative RT-PCR methodology (Gibson et al., this issue) has also been developed. Finally, real time quantitative PCR methodology can be used to develop high-throughput screening assays for a variety of applications [quantitative gene expression (RT-PCR), gene copy assays (Her2, HIV, etc.), genotyping (knockout mouse analysis), and immunoprecipitation].

Real-time PCR may also be performed using intercalating dyes (Higuchi et al. 1992) such as ethidium bromide. The fluorogenic probe method offers a major advantage over intercalating dyes—greater specificity (i.e., primer dimers and nonspecific PCR products are not detected).

- Hassler, H.A., S.J. Flood, K.J. Livak, J. Marmaro, R. Kohn, and C.A. Batt. 1995. Use of a fluorogenic probe in a PCR-based assay for the detection of *Listeria monocytogenes*. *App. Environ. Microbiol.* 61: 3724-3728.
- Hoecker-Andre, M. 1991. Quantitative evaluation of mRNA levels. *Meth. Mol. Cell. Biol.* 2: 189-201.
- Clement, M., S. Menzo, P. Bagnarelli, A. Manzù, A. Valenza, and P.E. Varaldo. 1993. Quantitative PCR and RT-PCR in virology. [Review]. *PCR Methods Applic.* 2: 191-196.
- Connor, R.J., H. Mofat, Y. Cao, and D.D. Ho. 1993. Increased viral burden and cytopathicity correlate temporally with CD4<sup>+</sup> T-lymphocyte decline and clinical progression in human immunodeficiency virus type 1-infected individuals. *J. Virol.* 67: 1772-1777.
- Eaton, D.L., W.J. Wood, D. Eaton, P.E. Hoss, P.

## HFID LI AL

Venar, and C. Gorman. 1986. Construction and characterization of an active factor VIII variant lacking the central one third of the molecule. *Biochemistry* 25: 8343-8347.

Fasco, M.J., C.P. Treanor, S. Spivack, H.L. Pigge, and L.S. Kaminsky. 1995. Quantitative RNA-polymerase chain reaction-DNA analysis by capillary electrophoresis and laser-induced fluorescence. *Anal. Biochem.* 224: 140-147.

Ferre, J. 1992. Quantitative or semi-quantitative PCR: Reality versus myth. *PCR Methods Applic.* 2: 1-9.

Furtado, M.R., L.A. Kingsley, and S.M. Wollinsky. 1995. Changes in the viral mRNA expression pattern correlate with a rapid rate of CD4<sup>+</sup> T-cell number decline in human immunodeficiency virus type 1-infected individuals. *J. Virol.* 69: 2092-2100.

Gibson, U.E.M., C.A. Heid, and P.M. Williams. 1996. A novel method for real time quantitative competitive RT-PCR. *Genome Res.* (this issue).

Gorman, C.M., D.R. Gies, and G. McCray. 1990. Transient production of proteins using an adenovirus transformed cell line. *DNA Prot. Engin. Tech.* 2: 3-10.

Higuchi, R., G. Dollinger, P.S. Walsh, and R. Griffith. 1992. Simultaneous amplification and detection of specific DNA sequences. *BioTechnology* 10: 413-417.

Holland, P.M., R.D. Abramson, R. Watson, and D.J. Gelfand. 1991. Detection of specific polymerase chain reaction product by utilizing the 5'-3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. Natl. Acad. Sci.* 88: 7276-7280.

Huang, S.K., H.Q. Xiao, T.J. Klein, G. Paciotti, D.G. Marsh, L.M. Lichtenstein, and M.C. Liu. 1995a. IL-13 expression at the sites of allergen challenge in patients with asthma. *J. Immun.* 155: 2688-2694.

Huang, S.K., M. Yi, E. Palmer, and D.G. Marsh. 1995b. A dominant T cell receptor beta-chain in response to a short ragweed allergen, Amb a 5. *J. Immun.* 154: 6157-6162.

Kellogg, D.E., J.J. Shinsky, and S. Kowk. 1990. Quantitation of HIV-1 proviral DNA relative to cellular DNA by the polymerase chain reaction. *Anal. Biochem.* 189: 202-208.

Lee, I.-G., C.R. Connell, and W. Bloch. 1993. Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Res.* 21: 3761-3766.

Livak, K.J., S.J. Flood, J. Marmaro, W. Gusti, and K. Dectz. 1995a. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nuclear acid hybridization. *PCR Methods Applic.* 4: 357-362.

Livak, K.J., J. Marmaro, and J.A. Todd. 1995b. Towards

fully automated genome-wide polymorphism screening [Letter]. *Nature Genet.* 9: 341-342.

Mulder, J., N. McKinney, C. Christopherson, J. Salsky, L. Greenfield, and S. Kwok. 1994. Rapid and simple PCR assay for quantitation of human immunodeficiency virus type 1 RNA in plasma: Application to acute retroviral infection. *J. Clin. Microbiol.* 32: 292-300.

Pang, S., Y. Koyanagi, S. Miles, C. Wiloy, H.V. Vinters, and L.S. Chen. 1990. High levels of unintegrated HIV-1 DNA in brain tissue of AIDS dementia patients. *Nature* 343: 85-89.

Platak, M.J., K.C. Luk, B. Williams, and J.D. Lifson. 1995a. Quantitative competitive polymerase chain reaction for accurate quantitation of HIV DNA and RNA species. *AltTechniques* 14: 70-81.

Platak, M.J., M.S. Saag, L.C. Yang, S.J. Clark, J.C. Kappes, K.C. Luk, B.H. Hann, G.M. Shaw, and J.D. Lifson. 1995b. High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR [see Comments]. *Science* 259: 1742-1754.

Prud'homme, G.J., D.H. Kono, and A.N. Theofilopoulos. 1995. Quantitative polymerase chain reaction analysis reveals marked overexpression of interleukin-1 beta, interleukin-1 and interferon-gamma mRNA in the lymph nodes of lupus-prone mice. *Mol. Immunol.* 32: 495-503.

Racymackers, L. 1995. A commentary on the practical applications of competitive PCR. *Genome Res.* 5: 91-94.

Sharp, P.A., A.J. Berk, and S.M. Herget. 1980. Transcription maps of adenovirus. *Methods Enzymol.* 65: 750-768.

Slamon, D.J., G.M. Clark, S.G. Wong, W.J. Levin, A. Ulrich, and W.J. McGuire. 1987. Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235: 177-182.

Southern, E.M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98: 503-517.

Tan, X., X. Sun, C.F. Gonzalez, and W. Hsueh. 1994. IAP and TNF increase the precursor of Nk-kappa B p50 mRNA in mouse intestine: Quantitative analysis by competitive PCR. *Biochim. Biophys. Acta* 1215: 157-162.

Thomas, P.S. 1980. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc. Natl. Acad. Sci.* 77: 5201-5205.

Williams, S., C. Schwer, A. Krishnasao, C. Heid, B. Karger, and P.M. Williams. 1996. Quantitative competitive PCR: Analysis of amplified products of the HIV-1 gag gene by capillary electrophoresis with laser induced fluorescence detection. *Anal. Biochem.* (in press).

Received June 3, 1996; accepted in revised form July 29, 1996.

## WISP genes are members of the connective tissue growth factor family that are up-regulated in Wnt-1-transformed cells and aberrantly expressed in human colon tumors

DIANE PENNICA\*†, TODD A. SWANSON\*, JAMES W. WELSH\*, MARGARET A. ROY‡, DAVID A. LAWRENCE\*, JAMES LEE‡, JENNIFER BRUSH‡, LISA A. TANEYHILL§, BETHANNE DEUEL‡, MICHAEL LEW¶, COLIN WATANABE||, ROBERT L. COHEN\*, MONA F. MELHEM\*\*, GENE G. FINLEY\*\*, PHIL QUIRKE††, AUDREY D. GODDARD‡, KENNETH J. HILLAN¶, AUSTIN L. GURNEY‡, DAVID BOTSTEIN†‡‡, AND ARNOLD J. LEVINE§

Departments of \*Molecular Oncology, †Molecular Biology, ‡Scientific Computing, and §Pathology, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080; \*\*University of Pittsburgh School of Medicine, Veterans Administration Medical Center, Pittsburgh, PA 15240; ††University of Leeds, Leeds, LS29JT United Kingdom; ‡‡Department of Genetics, Stanford University, Palo Alto, CA 94305; and ‡‡‡Department of Molecular Biology, Princeton University, Princeton, NJ 08544

Contributed by David Botstein and Arnold J. Levine, October 21, 1998

**ABSTRACT** Wnt family members are critical to many developmental processes, and components of the Wnt signaling pathway have been linked to tumorigenesis in familial and sporadic colon carcinomas. Here we report the identification of two genes, *WISP-1* and *WISP-2*, that are up-regulated in the mouse mammary epithelial cell line C57MG transformed by Wnt-1, but not by Wnt-4. Together with a third related gene, *WISP-3*, these proteins define a subfamily of the connective tissue growth factor family. Two distinct systems demonstrated *WISP* induction to be associated with the expression of Wnt-1. These included (i) C57MG cells infected with a Wnt-1 retroviral vector or expressing Wnt-1 under the control of a tetracycline repressible promoter, and (ii) Wnt-1 transgenic mice. The *WISP-1* gene was localized to human chromosome 8q24.1–8q24.3. *WISP-1* genomic DNA was amplified in colon cancer cell lines and in human colon tumors and its RNA overexpressed (2- to >30-fold) in 84% of the tumors examined compared with patient-matched normal mucosa. *WISP-3* mapped to chromosome 6q22–6q23 and also was overexpressed (4- to >40-fold) in 63% of the colon tumors analyzed. In contrast, *WISP-2* mapped to human chromosome 20q12–20q13 and its DNA was amplified, but RNA expression was reduced (2- to >30-fold) in 79% of the tumors. These results suggest that the *WISP* genes may be downstream of Wnt-1 signaling and that aberrant levels of *WISP* expression in colon cancer may play a role in colon tumorigenesis.

Wnt-1 is a member of an expanding family of cysteine-rich, glycosylated signaling proteins that mediate diverse developmental processes such as the control of cell proliferation, adhesion, cell polarity, and the establishment of cell fates (1, 2). Wnt-1 originally was identified as an oncogene activated by the insertion of mouse mammary tumor virus in virus-induced mammary adenocarcinomas (3, 4). Although Wnt-1 is not expressed in the normal mammary gland, expression of Wnt-1 in transgenic mice causes mammary tumors (5).

In mammalian cells, Wnt family members initiate signaling by binding to the seven-transmembrane spanning Frizzled receptors and recruiting the cytoplasmic protein Dishevelled (Dsh) to the cell membrane (1, 2, 6). Dsh then inhibits the kinase activity of the normally constitutively active glycogen synthase kinase-3 $\beta$  (GSK-3 $\beta$ ) resulting in an increase in  $\beta$ -catenin levels. Stabilized  $\beta$ -catenin interacts with the transcription factor TCF/Lef1, forming a complex that appears in

the nucleus and binds TCF/Lef1 target DNA elements to activate transcription (7, 8). Other experiments suggest that the adenomatous polyposis coli (APC) tumor suppressor gene also plays an important role in Wnt signaling by regulating  $\beta$ -catenin levels (9). APC is phosphorylated by GSK-3 $\beta$ , binds to  $\beta$ -catenin, and facilitates its degradation. Mutations in either APC or  $\beta$ -catenin have been associated with colon carcinomas and melanomas, suggesting these mutations contribute to the development of these types of cancer, implicating the Wnt pathway in tumorigenesis (1).

Although much has been learned about the Wnt signaling pathway over the past several years, only a few of the transcriptionally activated downstream components activated by Wnt have been characterized. Those that have been described cannot account for all of the diverse functions attributed to Wnt signaling. Among the candidate Wnt target genes are those encoding the nodal-related 3 gene, *Xnr3*, a member of the transforming growth factor (TGF)- $\beta$  superfamily, and the homeobox genes, *engrailed*, *goosecoid*, *twin* (*Xtwn*), and *siamois* (2). A recent report also identifies *c-myc* as a target gene of the Wnt signaling pathway (10).

To identify additional downstream genes in the Wnt signaling pathway that are relevant to the transformed cell phenotype, we used a PCR-based cDNA subtraction strategy, suppression subtractive hybridization (SSH) (11), using RNA isolated from C57MG mouse mammary epithelial cells and C57MG cells stably transformed by a Wnt-1 retrovirus. Overexpression of Wnt-1 in this cell line is sufficient to induce a partially transformed phenotype, characterized by elongated and refractile cells that lose contact inhibition and form a multilayered array (12, 13). We reasoned that genes differentially expressed between these two cell lines might contribute to the transformed phenotype.

In this paper, we describe the cloning and characterization of two genes up-regulated in Wnt-1 transformed cells, *WISP-1* and *WISP-2*, and a third related gene, *WISP-3*. The *WISP* genes are members of the CCN family of growth factors, which includes connective tissue growth factor (CTGF), Cyr61, and *nov*, a family not previously linked to Wnt signaling.

### MATERIALS AND METHODS

**SSH.** SSH was performed by using the PCR-Select cDNA Subtraction Kit (CLONTECH). Tester double-stranded

Abbreviations: TGF, transforming growth factor; CTGF, connective tissue growth factor; SSH, suppression subtractive hybridization; VWC, von Willebrand factor type C module.

Data deposition: The sequences reported in this paper have been deposited in the Genbank database (accession nos. AF100777, AF100778, AF100779, AF100780, and AF100781).

†To whom reprint requests should be addressed. e-mail: diane@gene.com.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/9514717-6\$2.00/0  
PNAS is available online at www.pnas.org.



cDNA was synthesized from 2  $\mu$ g of poly(A)<sup>+</sup> RNA isolated from the C57MG/Wnt-1 cell line and driver cDNA from 2  $\mu$ g of poly(A)<sup>+</sup> RNA from the parent C57MG cells. The subtracted cDNA library was subcloned into a pGEM-T vector for further analysis.

**cDNA Library Screening.** Clones encoding full-length mouse *WISP-1* were isolated by screening a  $\lambda$ gt10 mouse embryo cDNA library (CLONTECH) with a 70-bp probe from the original partial clone 568 sequence corresponding to amino acids 128–169. Clones encoding full-length human *WISP-1* were isolated by screening  $\lambda$ gt10 lung and fetal kidney cDNA libraries with the same probe at low stringency. Clones encoding full-length mouse and human *WISP-2* were isolated by screening a C57MG/Wnt-1 or human fetal lung cDNA library with a probe corresponding to nucleotides 1463–1512. Full-length cDNAs encoding *WISP-3* were cloned from human bone marrow and fetal kidney libraries.

**Expression of Human *WISP* RNA.** PCR amplification of first-strand cDNA was performed with human Multiple Tissue cDNA panels (CLONTECH) and 300  $\mu$ M of each dNTP at 94°C for 1 sec, 62°C for 30 sec, 72°C for 1 min, for 22–32 cycles. *WISP* and glyceraldehyde-3-phosphate dehydrogenase primer sequences are available on request.

**In Situ Hybridization.** <sup>32</sup>P-labeled sense and antisense riboprobes were transcribed from an 897-bp PCR product corresponding to nucleotides 601–1440 of mouse *WISP-1* or a 294-bp PCR product corresponding to nucleotides 82–375 of mouse *WISP-2*. All tissues were processed as described (40).

**Radiation Hybrid Mapping.** Genomic DNA from each hybrid in the Stanford G3 and Genebridge4 Radiation Hybrid Panels (Research Genetics, Huntsville, AL) and human and hamster control DNAs were PCR-amplified, and the results were submitted to the Stanford or Massachusetts Institute of Technology web servers.

**Cell Lines, Tumors, and Mucosa Specimens.** Tissue specimens were obtained from the Department of Pathology (University of Pittsburgh) for patients undergoing colon resection and from the University of Leeds, United Kingdom. Genomic DNA was isolated (Qiagen) from the pooled blood of 10 normal human donors, surgical specimens, and the following ATCC human cell lines: SW480, COLO 320DM, HT-29, WiDr, and SW403 (colon adenocarcinomas), SW620 (lymph node metastasis, colon adenocarcinoma), HCT 116 (colon carcinoma), SK-CO-1 (colon adenocarcinoma, ascites), and HM7 (a variant of ATCC colon adenocarcinoma cell line LS 174T). DNA concentration was determined by using Hoechst dye 33258 intercalation fluorimetry. Total RNA was prepared by homogenization in 7 M GuSCN followed by centrifugation over CsCl cushions or prepared by using RNeasy.

**Gene Amplification and RNA Expression Analysis.** Relative gene amplification and RNA expression of *WISPs* and *c-myc* in the cell lines, colorectal tumors, and normal mucosa were determined by quantitative PCR. Gene-specific primers and fluorogenic probes (sequences available on request) were designed and used to amplify and quantitate the genes. The relative gene copy number was derived by using the formula  $2^{-\Delta Ct}$  where  $\Delta Ct$  represents the difference in amplification cycles required to detect the *WISP* genes in peripheral blood lymphocyte DNA compared with colon tumor DNA or colon tumor RNA compared with normal mucosal RNA. The  $\delta$ -method was used for calculation of the SE of the gene copy number or RNA expression level. The *WISP*-specific signal was normalized to that of the glyceraldehyde-3-phosphate dehydrogenase housekeeping gene. All TaqMan assay reagents were obtained from Perkin-Elmer Applied Biosystems.

## RESULTS

**Isolation of *WISP-1* and *WISP-2* by SSH.** To identify Wnt-1-inducible genes, we used the technique of SSH using the

mouse mammary epithelial cell line C57MG and C57MG cells that stably express Wnt-1 (11). Candidate differentially expressed cDNAs (1,384 total) were sequenced. Thirty-nine percent of the sequences matched known genes or homologues, 32% matched expressed sequence tags, and 29% had no match. To confirm that the transcript was differentially expressed, semiquantitative reverse transcription-PCR and Northern analysis were performed by using mRNA from the C57MG and C57MG/Wnt-1 cells.

Two of the cDNAs, *WISP-1* and *WISP-2*, were differentially expressed, being induced in the C57MG/Wnt-1 cell line, but not in the parent C57MG cells or C57MG cells overexpressing Wnt-4 (Fig. 1 A and B). Wnt-4, unlike Wnt-1, does not induce the morphological transformation of C57MG cells and has no effect on  $\beta$ -catenin levels (13, 14). Expression of *WISP-1* was up-regulated approximately 3-fold in the C57MG/Wnt-1 cell line and *WISP-2* by approximately 5-fold by both Northern analysis and reverse transcription-PCR.

An independent, but similar, system was used to examine *WISP* expression after Wnt-1 induction. C57MG cells expressing the *Wnt-1* gene under the control of a tetracycline-repressible promoter produce low amounts of Wnt-1 in the repressed state but show a strong induction of *Wnt-1* mRNA and protein within 24 hr after tetracycline removal (8). The levels of Wnt-1 and *WISP* RNA isolated from these cells at various times after tetracycline removal were assessed by quantitative PCR. Strong induction of Wnt-1 mRNA was seen as early as 10 hr after tetracycline removal. Induction of *WISP* mRNA (2- to 6-fold) was seen at 48 and 72 hr (data not shown). These data support our previous observations that show that *WISP* induction is correlated with Wnt-1 expression. Because the induction is slow, occurring after approximately 48 hr, the induction of *WISPs* may be an indirect response to Wnt-1 signaling.

cDNA clones of human *WISP-1* were isolated and the sequence compared with mouse *WISP-1*. The cDNA sequences of mouse and human *WISP-1* were 1,766 and 2,830 bp in length, respectively, and encode proteins of 367 aa, with predicted relative molecular masses of  $\approx 40,000$  ( $M_r$ , 40 K). Both have hydrophobic N-terminal signal sequences, 38 conserved cysteine residues, and four potential N-linked glycosylation sites and are 84% identical (Fig. 2A).

Full-length cDNA clones of mouse and human *WISP-2* were 1,734 and 1,293 bp in length, respectively, and encode proteins of 251 and 250 aa, respectively, with predicted relative molecular masses of  $\approx 27,000$  ( $M_r$ , 27 K) (Fig. 2B). Mouse and human *WISP-2* are 73% identical. Human *WISP-2* has no potential N-linked glycosylation sites, and mouse *WISP-2* has one at

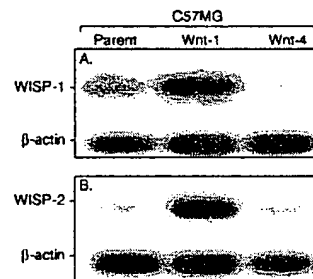


FIG. 1. *WISP-1* and *WISP-2* are induced by Wnt-1, but not Wnt-4, expression in C57MG cells. Northern analysis of *WISP-1* (A) and *WISP-2* (B) expression in C57MG, C57MG/Wnt-1, and C57MG/Wnt-4 cells. Poly(A)<sup>+</sup> RNA (2  $\mu$ g) was subjected to Northern blot analysis and hybridized with a 70-bp mouse *WISP-1*-specific probe (amino acids 278–300) or a 190-bp *WISP-2*-specific probe (nucleotides 1438–1627) in the 3' untranslated region. Blots were rehybridized with human  $\beta$ -actin probe.

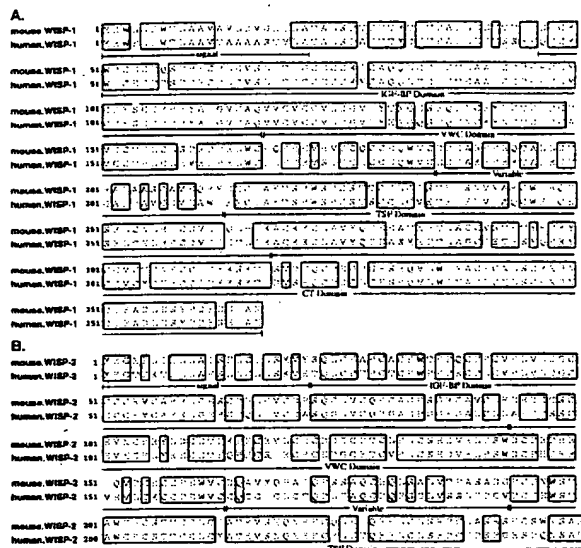


FIG. 2. Encoded amino acid sequence alignment of mouse and human *WISP-1* (A) and mouse and human *WISP-2* (B). The potential signal sequence, insulin-like growth factor-binding protein (IGF-BP), VWC, thrombospondin (TSP), and C-terminal (CT) domains are underlined.

position 197. *WISP-2* has 28 cysteine residues that are conserved among the 38 cysteines found in *WISP-1*.

**Identification of *WISP-3*.** To search for related proteins, we screened expressed sequence tag (EST) databases with the *WISP-1* protein sequence and identified several ESTs as potentially related sequences. We identified a homologous protein that we have called *WISP-3*. A full-length human *WISP-3* cDNA of 1,371 bp was isolated corresponding to those ESTs that encode a 354-aa protein with a predicted molecular mass of 39,293. *WISP-3* has two potential N-linked glycosylation sites and 36 cysteine residues. An alignment of the three human *WISP* proteins shows that *WISP-1* and *WISP-3* are the most similar (42% identity), whereas *WISP-2* has 37% identity with *WISP-1* and 32% identity with *WISP-3* (Fig. 3A).

***WISPs* Are Homologous to the CTGF Family of Proteins.** Human *WISP-1*, *WISP-2*, and *WISP-3* are novel sequences; however, mouse *WISP-1* is the same as the recently identified *Elm1* gene. *Elm1* is expressed in low, but not high, metastatic mouse melanoma cells, and suppresses the *in vivo* growth and metastatic potential of K-1735 mouse melanoma cells (15). Human and mouse *WISP-2* are homologous to the recently described rat gene, *rCop-1* (16). Significant homology (36–44%) was seen to the CCN family of growth factors. This family includes three members, CTGF, Cyr61, and the protooncogene *nov*. CTGF is a chemotactic and mitogenic factor for fibroblasts that is implicated in wound healing and fibrotic disorders and is induced by TGF- $\beta$  (17). Cyr61 is an extracellular matrix signaling molecule that promotes cell adhesion, proliferation, migration, angiogenesis, and tumor growth (18, 19). *nov* (nephroblastoma overexpressed) is an immediate early gene associated with quiescence and found altered in Wilms tumors (20). The proteins of the CCN family share functional, but not sequence, similarity to Wnt-1. All are secreted, cysteine-rich heparin binding glycoproteins that associate with the cell surface and extracellular matrix.

*WISP* proteins exhibit the modular architecture of the CCN family, characterized by four conserved cysteine-rich domains (Fig. 3B) (21). The N-terminal domain, which includes the first 12 cysteine residues, contains a consensus sequence (GCGC-CXXC) conserved in most insulin-like growth factor (IGF)-

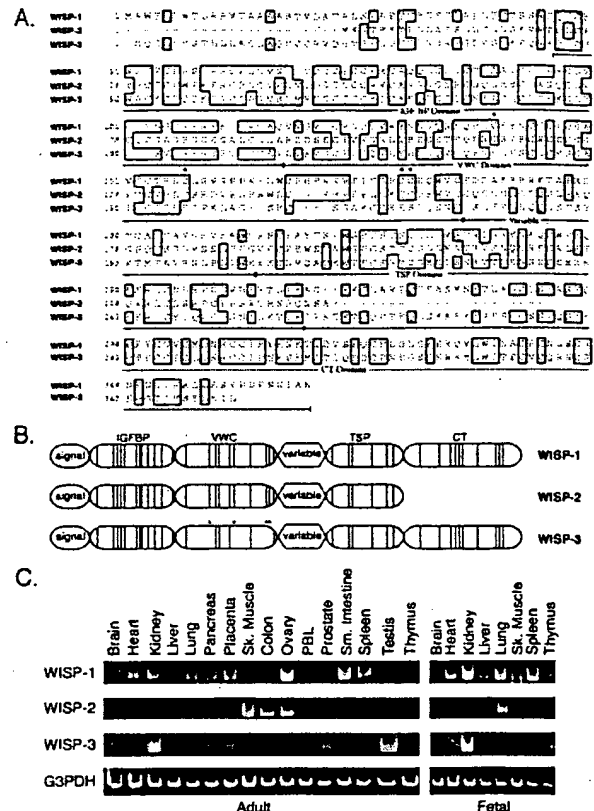


FIG. 3. (A) Encoded amino acid sequence alignment of human *WISPs*. The cysteine residues of *WISP-1* and *WISP-2* that are not present in *WISP-3* are indicated with a dot. (B) Schematic representation of the *WISP* proteins showing the domain structure and cysteine residues (vertical lines). The four cysteine residues in the VWC domain that are absent in *WISP-3* are indicated with a dot. (C) Expression of *WISP* mRNA in human tissues. PCR was performed on human multiple-tissue cDNA panels (CLONTECH) from the indicated adult and fetal tissues.

binding proteins (BP). This sequence is conserved in *WISP-2* and *WISP-3*, whereas *WISP-1* has a glutamine in the third position instead of a glycine. CTGF recently has been shown to specifically bind IGF (22) and a truncated *nov* protein lacking the IGF-BP domain is oncogenic (23). The von Willebrand factor type C module (VWC), also found in certain collagens and mucins, covers the next 10 cysteine residues, and is thought to participate in protein complex formation and oligomerization (24). The VWC domain of *WISP-3* differs from all CCN family members described previously, in that it contains only six of the 10 cysteine residues (Fig. 3A and B). A short variable region follows the VWC domain. The third module, the thrombospondin (TSP) domain is involved in binding to sulfated glycoconjugates and contains six cysteine residues and a conserved WSxCSxxCG motif first identified in thrombospondin (25). The C-terminal (CT) module containing the remaining 10 cysteines is thought to be involved in dimerization and receptor binding (26). The CT domain is present in all CCN family members described to date but is absent in *WISP-2* (Fig. 3A and B). The existence of a putative signal sequence and the absence of a transmembrane domain suggest that *WISPs* are secreted proteins, an observation supported by an analysis of their expression and secretion from mammalian cell and baculovirus cultures (data not shown).

**Expression of *WISP* mRNA in Human Tissues.** Tissue-specific expression of human *WISPs* was characterized by PCR

analysis on adult and fetal multiple tissue cDNA panels. *WISP-1* expression was seen in the adult heart, kidney, lung, pancreas, placenta, ovary, small intestine, and spleen (Fig. 3C). Little or no expression was detected in the brain, liver, skeletal muscle, colon, peripheral blood leukocytes, prostate, testis, or thymus. *WISP-2* had a more restricted tissue expression and was detected in adult skeletal muscle, colon, ovary, and fetal lung. Predominant expression of *WISP-3* was seen in adult kidney and testis and fetal kidney. Lower levels of *WISP-3* expression were detected in placenta, ovary, prostate, and small intestine.

**In Situ Localization of *WISP-1* and *WISP-2*.** Expression of *WISP-1* and *WISP-2* was assessed by *in situ* hybridization in mammary tumors from Wnt-1 transgenic mice. Strong expression of *WISP-1* was observed in stromal fibroblasts lying within the fibrovascular tumor stroma (Fig. 4 A–D). However, low-level *WISP-1* expression also was observed focally within tumor cells (data not shown). No expression was observed in normal breast. Like *WISP-1*, *WISP-2* expression also was seen in the tumor stroma in breast tumors from Wnt-1 transgenic animals (Fig. 4 E–H). However, *WISP-2* expression in the stroma was in spindle-shaped cells adjacent to capillary vessels, whereas

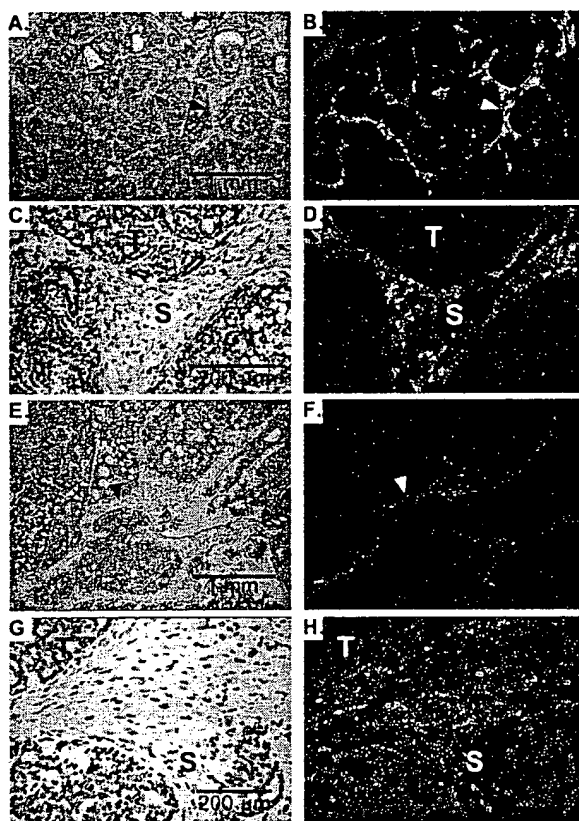


FIG. 4. (A, C, E, and G) Representative hematoxylin/eosin-stained images from breast tumors in Wnt-1 transgenic mice. The corresponding dark-field images showing *WISP-1* expression are shown in B and D. The tumor is a moderately well-differentiated adenocarcinoma showing evidence of adenoid cystic change. At low power (A and B), expression of *WISP-1* is seen in the delicate branching fibrovascular tumor stroma (arrowhead). At higher magnification, expression is seen in the stromal(s) fibroblasts (C and D), and tumor cells are negative. Focal expression of *WISP-1*, however, was observed in tumor cells in some areas. Images of *WISP-2* expression are shown in E–H. At low power (E and F), expression of *WISP-2* is seen in cells lying within the fibrovascular tumor stroma. At higher magnification, these cells appeared to be adjacent to capillary vessels whereas tumor cells are negative (G and H).

the predominant cell type expressing *WISP-1* was the stromal fibroblasts.

**Chromosome Localization of the *WISP* Genes.** The chromosomal location of the human *WISP* genes was determined by radiation hybrid mapping panels. *WISP-1* is approximately 3.48 cR from the meiotic marker AFM259xc5 [logarithm of odds (lod) score 16.31] on chromosome 8q24.1 to 8q24.3, in the same region as the human locus of the *novH* family member (27) and roughly 4 Mbs distal to *c-myc* (28). Preliminary fine mapping indicates that *WISP-1* is located near D8S1712 STS. *WISP-2* is linked to the marker SHGC-33922 (lod = 1,000) on chromosome 20q12–20q13.1. Human *WISP-3* mapped to chromosome 6q22–6q23 and is linked to the marker AFM211ze5 (lod = 1,000). *WISP-3* is approximately 18 Mbs proximal to CTGF and 23 Mbs proximal to the human cellular oncogene *MYB* (27, 29).

**Amplification and Aberrant Expression of *WISPs* in Human Colon Tumors.** Amplification of protooncogenes is seen in many human tumors and has etiological and prognostic significance. For example, in a variety of tumor types, *c-myc* amplification has been associated with malignant progression and poor prognosis (30). Because *WISP-1* resides in the same general chromosomal location (8q24) as *c-myc*, we asked whether it was a target of gene amplification, and, if so, whether this amplification was independent of the *c-myc* locus. Genomic DNA from human colon cancer cell lines was assessed by quantitative PCR and Southern blot analysis (Fig. 5 A and B). Both methods detected similar degrees of *WISP-1* amplification. Most cell lines showed significant (2- to 4-fold) amplification, with the HT-29 and WiDr cell lines demonstrating an 8-fold increase. Significantly, the pattern of amplification observed did not correlate with that observed for *c-myc*, indicating that the *c-myc* gene is not part of the amplicon that involves the *WISP-1* locus.

We next examined whether the *WISP* genes were amplified in a panel of 25 primary human colon adenocarcinomas. The relative *WISP* gene copy number in each colon tumor DNA was compared with pooled normal DNA from 10 donors by quantitative PCR (Fig. 6). The copy number of *WISP-1* and *WISP-2* was significantly greater than one, approximately 2-fold for *WISP-1* in about 60% of the tumors and 2- to 4-fold for *WISP-2* in 92% of the tumors ( $P < 0.001$  for each). The copy number for *WISP-3* was indistinguishable from one ( $P = 0.166$ ). In addition, the copy number of *WISP-2* was significantly higher than that of *WISP-1* ( $P < 0.001$ ).

The levels of *WISP* transcripts in RNA isolated from 19 adenocarcinomas and their matched normal mucosa were

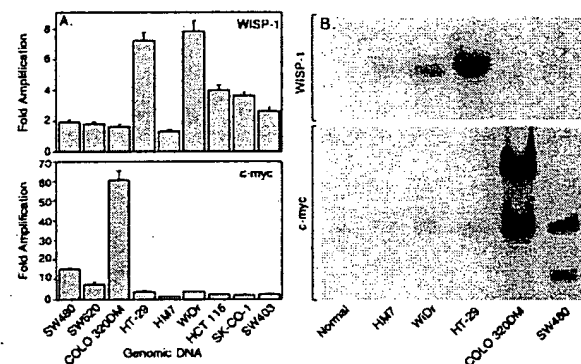


FIG. 5. Amplification of *WISP-1* genomic DNA in colon cancer cell lines. (A) Amplification in cell line DNA was determined by quantitative PCR. (B) Southern blots containing genomic DNA (10  $\mu$ g) digested with *EcoRI* (*WISP-1*) or *XbaI* (*c-myc*) were hybridized with a 100-bp human *WISP-1* probe (amino acids 186–219) or a human *c-myc* probe (located at bp 1901–2000). The *WISP* and *myc* genes are detected in normal human genomic DNA after a longer film exposure.

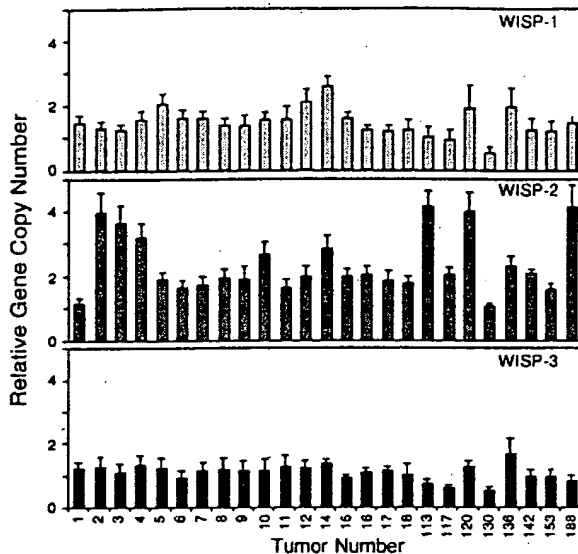


FIG. 6. Genomic amplification of *WISP* genes in human colon tumors. The relative gene copy number of the *WISP* genes in 25 adenocarcinomas was assayed by quantitative PCR, by comparing DNA from primary human tumors with pooled DNA from 10 healthy donors. The data are means  $\pm$  SEM from one experiment done in triplicate. The experiment was repeated at least three times.

assessed by quantitative PCR (Fig. 7). The level of *WISP-1* RNA present in tumor tissue varied but was significantly increased (2- to >25-fold) in 84% (16/19) of the human colon tumors examined compared with normal adjacent mucosa. Four of 19 tumors showed greater than 10-fold overexpression. In contrast, in 79% (15/19) of the tumors examined, *WISP-2* RNA expression was significantly lower in the tumor than the mucosa. Similar to *WISP-1*, *WISP-3* RNA was overexpressed in 63% (12/19) of the colon tumors compared with the normal

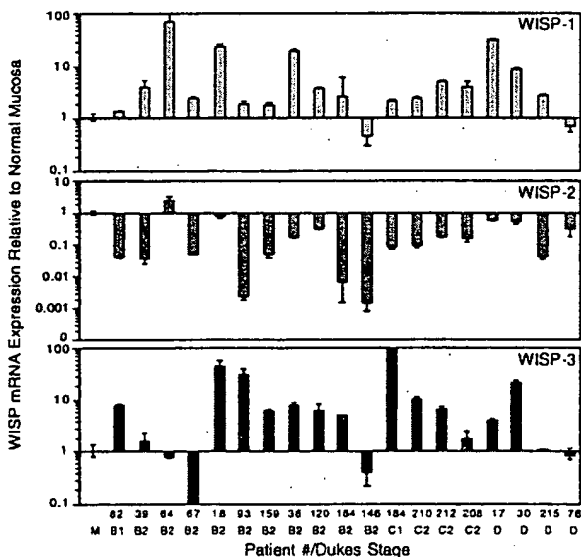


FIG. 7. *WISP* RNA expression in primary human colon tumors relative to expression in normal mucosa from the same patient. Expression of *WISP* mRNA in 19 adenocarcinomas was assayed by quantitative PCR. The Dukes stage of the tumor is listed under the sample number. The data are means  $\pm$  SEM from one experiment done in triplicate. The experiment was repeated at least twice.

mucosa. The amount of overexpression of *WISP-3* ranged from 4- to >40-fold.

## DISCUSSION

One approach to understanding the molecular basis of cancer is to identify differences in gene expression between cancer cells and normal cells. Strategies based on assumptions that steady-state mRNA levels will differ between normal and malignant cells have been used to clone differentially expressed genes (31). We have used a PCR-based selection strategy, SSH, to identify genes selectively expressed in C57MG mouse mammary epithelial cells transformed by Wnt-1.

Three of the genes isolated, *WISP-1*, *WISP-2*, and *WISP-3*, are members of the CCN family of growth factors, which includes CTGF, Cyr61, and *nov*, a family not previously linked to Wnt signaling.

Two independent experimental systems demonstrated that *WISP* induction was associated with the expression of Wnt-1. The first was C57MG cells infected with a Wnt-1 retroviral vector or C57MG cells expressing Wnt-1 under the control of a tetracycline-repressible promoter, and the second was in Wnt-1 transgenic mice, where breast tissue expresses Wnt-1, whereas normal breast tissue does not. No *WISP* RNA expression was detected in mammary tumors induced by polyoma virus middle T antigen (data not shown). These data suggest a link between Wnt-1 and *WISPs* in that in these two situations, *WISP* induction was correlated with Wnt-1 expression.

It is not clear whether the *WISPs* are directly or indirectly induced by the downstream components of the Wnt-1 signaling pathway (i.e.,  $\beta$ -catenin-TCF-1/Lef1). The increased levels of *WISP* RNA were measured in Wnt-1-transformed cells, hours or days after Wnt-1 transformation. Thus, *WISP* expression could result from Wnt-1 signaling directly through  $\beta$ -catenin transcription factor regulation or alternatively through Wnt-1 signaling turning on a transcription factor, which in turn regulates *WISPs*.

The *WISPs* define an additional subfamily of the CCN family of growth factors. One striking difference observed in the protein sequence of *WISP-2* is the absence of a CT domain, which is present in CTGF, Cyr61, *nov*, *WISP-1*, and *WISP-3*. This domain is thought to be involved in receptor binding and dimerization. Growth factors, such as TGF- $\beta$ , platelet-derived growth factor, and nerve growth factor, which contain a cysteine knot motif exist as dimers (32). It is tempting to speculate that *WISP-1* and *WISP-3* may exist as dimers, whereas *WISP-2* exists as a monomer. If the CT domain is also important for receptor binding, *WISP-2* may bind its receptor through a different region of the molecule than the other CCN family members. No specific receptors have been identified for CTGF or *nov*. A recent report has shown that integrin  $\alpha_v\beta_3$  serves as an adhesion receptor for Cyr61 (33).

The strong expression of *WISP-1* and *WISP-2* in cells lying within the fibrovascular tumor stroma in breast tumors from Wnt-1 transgenic animals is consistent with previous observations that transcripts for the related CTGF gene are primarily expressed in the fibrous stroma of mammary tumors (34). Epithelial cells are thought to control the proliferation of connective tissue stroma in mammary tumors by a cascade of growth factor signals similar to that controlling connective tissue formation during wound repair. It has been proposed that mammary tumor cells or inflammatory cells at the tumor interstitial interface secrete TGF- $\beta$ 1, which is the stimulus for stromal proliferation (34). TGF- $\beta$ 1 is secreted by a large percentage of malignant breast tumors and may be one of the growth factors that stimulates the production of CTGF and *WISPs* in the stroma.

It was of interest that *WISP-1* and *WISP-2* expression was observed in the stromal cells that surrounded the tumor cells

(epithelial cells) in the Wnt-1 transgenic mouse sections of breast tissue. This finding suggests that paracrine signaling could occur in which the stromal cells could supply WISP-1 and WISP-2 to regulate tumor cell growth on the WISP extracellular matrix. Stromal cell-derived factors in the extracellular matrix have been postulated to play a role in tumor cell migration and proliferation (35). The localization of *WISP-1* and *WISP-2* in the stromal cells of breast tumors supports this paracrine model.

An analysis of *WISP-1* gene amplification and expression in human colon tumors showed a correlation between DNA amplification and overexpression, whereas overexpression of *WISP-3* RNA was seen in the absence of DNA amplification. In contrast, *WISP-2* DNA was amplified in the colon tumors, but its mRNA expression was significantly reduced in the majority of tumors compared with the expression in normal colonic mucosa from the same patient. The gene for human *WISP-2* was localized to chromosome 20q12–20q13, at a region frequently amplified and associated with poor prognosis in node negative breast cancer and many colon cancers, suggesting the existence of one or more oncogenes at this locus (36–38). Because the center of the 20q13 amplicon has not yet been identified, it is possible that the apparent amplification observed for *WISP-2* may be caused by another gene in this amplicon.

A recent manuscript on *rCop-1*, the rat orthologue of *WISP-2*, describes the loss of expression of this gene after cell transformation, suggesting it may be a negative regulator of growth in cell lines (16). Although the mechanism by which *WISP-2* RNA expression is down-regulated during malignant transformation is unknown, the reduced expression of *WISP-2* in colon tumors and cell lines suggests that it may function as a tumor suppressor. These results show that the *WISP* genes are aberrantly expressed in colon cancer and suggest that their altered expression may confer selective growth advantage to the tumor.

Members of the Wnt signaling pathway have been implicated in the pathogenesis of colon cancer, breast cancer, and melanoma, including the tumor suppressor gene adenomatous polyposis coli and  $\beta$ -catenin (39). Mutations in specific regions of either gene can cause the stabilization and accumulation of cytoplasmic  $\beta$ -catenin, which presumably contributes to human carcinogenesis through the activation of target genes such as the *WISPs*. Although the mechanism by which Wnt-1 transforms cells and induces tumorigenesis is unknown, the identification of *WISPs* as genes that may be regulated downstream of Wnt-1 in C57MG cells suggests they could be important mediators of Wnt-1 transformation. The amplification and altered expression patterns of the *WISPs* in human colon tumors may indicate an important role for these genes in tumor development.

We thank the DNA synthesis group for oligonucleotide synthesis, T. Baker for technical assistance, P. Dowd for radiation hybrid mapping, K. Willert and R. Nusse for the tet-repressible C57MG/Wnt-1 cells, V. Dixit for discussions, and D. Wood and A. Bruce for artwork.

- Cadigan, K. M. & Nusse, R. (1997) *Genes Dev.* 11, 3286–3305.
- Dale, T. C. (1998) *Biochem. J.* 329, 209–223.
- Nusse, R. & Varmus, H. E. (1982) *Cell* 31, 99–109.
- van Ooyen, A. & Nusse, R. (1984) *Cell* 39, 233–240.
- Tsukamoto, A. S., Grosschedl, R., Guzman, R. C., Parslow, T. & Varmus, H. E. (1988) *Cell* 55, 619–625.
- Brown, J. D. & Moon, R. T. (1998) *Curr. Opin. Cell Biol.* 10, 182–187.
- Molenaar, M., van de Wetering, M., Oosterwegel, M., Peterson-Maduro, J., Godsave, S., Korinek, V., Roose, J., Destree, O. & Clevers, H. (1996) *Cell* 86, 391–399.
- Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O. & Clevers, H. (1998) *Mol. Cell Biol.* 18, 1248–1256.
- Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B. & Polakis, P. (1995) *Proc. Natl. Acad. Sci. USA* 92, 3046–3050.
- He, T. C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B. & Kinzler, K. W. (1998) *Science* 281, 1509–1512.
- Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. & Siebert, P. D. (1996) *Proc. Natl. Acad. Sci. USA* 93, 6025–6030.
- Brown, A. M., Wildin, R. S., Prendergast, T. J. & Varmus, H. E. (1986) *Cell* 46, 1001–1009.
- Wong, G. T., Gavin, B. J. & McMahon, A. P. (1994) *Mol. Cell Biol.* 14, 6278–6286.
- Shimizu, H., Julius, M. A., Giarre, M., Zheng, Z., Brown, A. M. & Kitajewski, J. (1997) *Cell Growth Differ.* 8, 1349–1358.
- Hashimoto, Y., Shindo-Okada, N., Tani, M., Nagamachi, Y., Takeuchi, K., Shiroishi, T., Toma, H. & Yokota, J. (1998) *J. Exp. Med.* 187, 289–296.
- Zhang, R., Averboukh, L., Zhu, W., Zhang, H., Jo, H., Dempsey, P. J., Coffey, R. J., Pardee, A. B. & Liang, P. (1998) *Mol. Cell Biol.* 18, 6131–6141.
- Grotendorst, G. R. (1997) *Cytokine Growth Factor Rev.* 8, 171–179.
- Kireeva, M. L., Mo, F. E., Yang, G. P. & Lau, L. F. (1996) *Mol. Cell Biol.* 16, 1326–1334.
- Babic, A. M., Kireeva, M. L., Kolesnikova, T. V. & Lau, L. F. (1998) *Proc. Natl. Acad. Sci. USA* 95, 6355–6360.
- Martinerie, C., Huff, V., Joubert, I., Badzioch, M., Saunders, G., Strong, L. & Perbal, B. (1994) *Oncogene* 9, 2729–2732.
- Bork, P. (1993) *FEBS Lett.* 327, 125–130.
- Kim, H. S., Nagalla, S. R., Oh, Y., Wilson, E., Roberts, C. T., Jr. & Rosenfeld, R. G. (1997) *Proc. Natl. Acad. Sci. USA* 94, 12981–12986.
- Joliot, V., Martinerie, C., Dambrine, G., Plassiart, G., Brisac, M., Crochet, J. & Perbal, B. (1992) *Mol. Cell Biol.* 12, 10–21.
- Mancuso, D. J., Tuley, E. A., Westfield, L. A., Worrall, N. K., Shelton-Inloes, B. B., Sorace, J. M., Alevy, Y. G. & Sadler, J. E. (1989) *J. Biol. Chem.* 264, 19514–19527.
- Holt, G. D., Pangburn, M. K. & Ginsburg, V. (1990) *J. Biol. Chem.* 265, 2852–2855.
- Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., van Mourik, J. A. & Pannekoek, H. (1991) *J. Cell Biol.* 113, 195–205.
- Martinerie, C., Viegas-Pequignot, E., Guenard, I., Dutrillaux, B., Nguyen, V. C., Bernheim, A. & Perbal, B. (1992) *Oncogene* 7, 2529–2534.
- Takahashi, E., Hori, T., O'Connell, P., Leppert, M. & White, R. (1991) *Cytogenet. Cell Genet.* 57, 109–111.
- Meese, E., Meltzer, P. S., Witkowski, C. M. & Trent, J. M. (1989) *Genes Chromosomes Cancer* 1, 88–94.
- Garte, S. J. (1993) *Crit. Rev. Oncog.* 4, 435–449.
- Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997) *Science* 276, 1268–1272.
- Sun, P. D. & Davies, D. R. (1995) *Annu. Rev. Biophys. Biomol. Struct.* 24, 269–291.
- Kireeva, M. L., Lam, S. C. T. & Lau, L. F. (1998) *J. Biol. Chem.* 273, 3090–3096.
- Frazier, K. S. & Grotendorst, G. R. (1997) *Int. J. Biochem. Cell Biol.* 29, 153–161.
- Wernert, N. (1997) *Virchows Arch.* 430, 433–443.
- Tanner, M. M., Tirkkonen, M., Kallioniemi, A., Collins, C., Stokke, T., Karhu, R., Kowbel, D., Shadravan, F., Hintz, M., Kuo, W. L., *et al.* (1994) *Cancer Res.* 54, 4257–4260.
- Brinkmann, U., Gallo, M., Polymeropoulos, M. H. & Pastan, I. (1996) *Genome Res.* 6, 187–194.
- Bischoff, J. R., Anderson, L., Zhu, Y., Mossie, K., Ng, L., Souza, B., Schryver, B., Flanagan, P., Clairvoyant, F., Ginther, C., *et al.* (1998) *EMBO J.* 17, 3052–3065.
- Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B. & Kinzler, K. W. (1997) *Science* 275, 1787–1790.
- Lu, L. H. & Gillett, N. (1994) *Cell Vision* 1, 169–176.

methods. Peptides AENK or AEQK were dissolved in water, made isotonic with NaCl and diluted into RPMI growth medium. T-cell-proliferation assays were done essentially as described<sup>20,21</sup>. Briefly, after antigen pulsing ( $30 \mu\text{g ml}^{-1}$  TTCF) with tetrapeptides ( $1\text{--}2 \text{ mg ml}^{-1}$ ), PBMCs or EBV-B cells were washed in PBS and fixed for 45 s in 0.05% glutaraldehyde. Glycine was added to a final concentration of 0.1M and the cells were washed five times in RPMI 1640 medium containing 1% FCS before co-culture with T-cell clones in round-bottom 96-well microtitre plates. After 48 h, the cultures were pulsed with  $1 \mu\text{Ci}$  of  $^3\text{H}$ -thymidine and harvested for scintillation counting 16 h later. Predigestion of native TTCF was done by incubating  $200 \mu\text{g}$  TTCF with  $0.25 \mu\text{g}$  pig kidney legumain in  $500 \mu\text{l}$  50 mM citrate buffer, pH 5.5, for 1 h at  $37^\circ\text{C}$ . **Glycopeptide digestions.** The peptides HIDNEEDI, HIDN(N-glucosamine) EEDI and HIDNESDI, which are based on the TTCF sequence, and QQQHIFGSGNVTDCSGNFCLFR(KKK), which is based on human transferrin, were obtained by custom synthesis. The three C-terminal lysine residues were added to the natural sequence to aid solubility. The transferrin glycopeptide QQQHIFGSGNVTDCSGNFCLFR was prepared by tryptic (Promega) digestion of 5 mg reduced, carboxy-methylated human transferrin followed by concanavalin A chromatography<sup>11</sup>. Glycopeptides corresponding to residues 622–642 and 421–452 were isolated by reverse-phase HPLC and identified by mass spectrometry and N-terminal sequencing. The lyophilized transferrin-derived peptides were redissolved in 50 mM sodium acetate, pH 5.5, 10 mM dithiothreitol, 20% methanol. Digestions were performed for 3 h at  $30^\circ\text{C}$  with  $5\text{--}50 \text{ mU ml}^{-1}$  pig kidney legumain or B-cell AEP. Products were analysed by HPLC or MALDI-TOF mass spectrometry using a matrix of  $10 \text{ mg ml}^{-1}$   $\alpha$ -cyanocinnamic acid in 50% acetonitrile/0.1% TFA and a PerSeptive Biosystems Elite STR mass spectrometer set to linear or reflector mode. Internal standardization was obtained with a matrix ion of 568.13 mass units.

Received 29 September; accepted 3 November 1998.

- Chen, J. M. *et al.* Cloning, isolation, and characterisation of mammalian legumain, an asparaginyl endopeptidase. *J. Biol. Chem.* 272, 8090–8098 (1997).
- Kembhavi, A. A., Buttle, D. J., Knight, C. G. & Barrett, A. J. The two cysteine endopeptidases of legume seeds: purification and characterization by use of specific fluorometric assays. *Arch. Biochem. Biophys.* 303, 208–213 (1993).
- Dalton, J. P., Hala Jambrikska, L. & Bridley, P. J. Asparaginyl endopeptidase activity in adult *Schistosoma mansoni*. *Parasitology* 111, 575–580 (1995).
- Bennett, K. *et al.* Antigen processing for presentation by class II major histocompatibility complex requires cleavage by cathepsin E. *Eur. J. Immunol.* 22, 1519–1524 (1992).
- Riese, R. J. *et al.* Essential role for cathepsin S in MHC class II-associated invariant chain processing and peptide loading. *Immunity* 4, 357–366 (1996).
- Rodríguez, G. M. & Diment, S. Role of cathepsin D in antigen presentation of ovalbumin. *J. Immunol.* 149, 2894–2898 (1992).
- Hewitt, E. W. *et al.* Natural processing sites for human cathepsin E and cathepsin D in tetanus toxin: implications for T cell epitope generation. *J. Immunol.* 159, 4693–4699 (1997).
- Watts, C. Capture and processing of exogenous antigens for presentation on MHC molecules. *Annu. Rev. Immunol.* 15, 821–850 (1997).
- Chapman, H. A. Endosomal proteases and MHC class II function. *Curr. Opin. Immunol.* 10, 93–102 (1998).
- Fineschi, B. & Miller, J. Endosomal proteases and antigen processing. *Trends Biochem. Sci.* 22, 377–382 (1997).
- Lu, J. & van Halbeek, H. Complete  $^1\text{H}$  and  $^{13}\text{C}$  resonance assignments of a 21-amino acid glycopeptide prepared from human serum transferrin. *Carbohydr. Res.* 296, 1–21 (1996).
- Fearon, D. T. & Locksley, R. M. The instructive role of innate immunity in the acquired immune response. *Science* 272, 50–54 (1996).
- Medzhitov, R. & Janeway, C. A. J. Innate immunity: the virtues of a nonclonal system of recognition. *Cell* 91, 295–298 (1997).
- Wyatt, R. *et al.* The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* 393, 705–711 (1998).
- Botarelli, P. *et al.* N-glycosylation of HIV gp120 may constrain recognition by T lymphocytes. *J. Immunol.* 147, 3128–3132 (1991).
- Davidson, H. W., West, M. A. & Watts, C. Endocytosis, intracellular trafficking, and processing of membrane IgG and monovalent antigen/membrane IgG complexes in B lymphocytes. *J. Immunol.* 144, 4101–4109 (1990).
- Barrett, A. J. & Kirschke, H. Cathepsin B, cathepsin H and cathepsin L. *Methods Enzymol.* 80, 535–559 (1981).
- Makoff, A. J., Ballantine, S. P., Smallwood, A. E. & Fairweather, N. F. Expression of tetanus toxin fragment C in *E. coli*: its purification and potential use as a vaccine. *Biotechnology* 7, 1043–1046 (1989).
- Lanc, D. P. & Harlow, E. *Antibodies: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, 1988).
- Lanzavecchia, A. Antigen-specific interaction between T and B cells. *Nature* 314, 537–539 (1985).
- Pond, L. & Watts, C. Characterization of transport of newly assembled, T cell-stimulatory MHC class II-peptide complexes from MHC class II compartments to the cell surface. *J. Immunol.* 159, 543–553 (1997).

**Acknowledgements.** We thank M. Ferguson for helpful discussions and advice; E. Smythe and L. Grayson for advice and technical assistance; B. Spruce, A. Knight and the BTS (Ninewells Hospital) for help with blood monocyte preparation; and our colleagues for many helpful comments on the manuscript. This work was supported by the Wellcome Trust and by an EMBO Long-term fellowship to B. M.

Correspondence and requests for materials should be addressed to C.W. (e-mail: c.watts@dundee.ac.uk).

## Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer

Robert M. Pitti<sup>††</sup>, Scot A. Marsters<sup>††</sup>, David A. Lawrence<sup>††</sup>, Margaret Roy<sup>\*</sup>, Frank C. Kischkel<sup>\*</sup>, Patrick Dowd<sup>\*</sup>, Arthur Huang<sup>\*</sup>, Christopher J. Donahue<sup>\*</sup>, Steven W. Sherwood<sup>\*</sup>, Daryl T. Baldwin<sup>\*</sup>, Paul J. Godowski<sup>\*</sup>, William I. Wood<sup>\*</sup>, Austin L. Gurney<sup>\*</sup>, Kenneth J. Hillan<sup>\*</sup>, Robert L. Cohen<sup>\*</sup>, Audrey D. Goddard<sup>\*</sup>, David Botstein<sup>‡</sup> & Avi Ashkenazi<sup>\*</sup>

<sup>\*</sup> Departments of Molecular Oncology, Molecular Biology, and Immunology, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA

<sup>‡</sup> Department of Genetics, Stanford University, Stanford, California 94305, USA

<sup>†</sup> These authors contributed equally to this work

Fas ligand (FasL) is produced by activated T cells and natural killer cells and it induces apoptosis (programmed cell death) in target cells through the death receptor Fas/Apo1/CD95 (ref. 1). One important role of FasL and Fas is to mediate immune-cytotoxic killing of cells that are potentially harmful to the organism, such as virus-infected or tumour cells<sup>1</sup>. Here we report the discovery of a soluble decoy receptor, termed decoy receptor 3 (Dcr3), that binds to FasL and inhibits FasL-induced apoptosis. The Dcr3 gene was amplified in about half of 35 primary lung and colon tumours studied, and Dcr3 messenger RNA was expressed in malignant tissue. Thus, certain tumours may escape FasL-dependent immune-cytotoxic attack by expressing a decoy receptor that blocks FasL.

By searching expressed sequence tag (EST) databases, we identified a set of related ESTs that showed homology to the tumour necrosis factor (TNF) receptor (TNFR) gene superfamily<sup>2</sup>. Using the overlapping sequence, we isolated a previously unknown full-length complementary DNA from human fetal lung. We named the protein encoded by this cDNA decoy receptor 3 (Dcr3). The cDNA encodes a 300-amino-acid polypeptide that resembles members of the TNFR family (Fig. 1a): the amino terminus contains a leader sequence, which is followed by four tandem cysteine-rich domains (CRDs). Like one other TNFR homologue, osteoprotegerin (OPG)<sup>3</sup>, Dcr3 lacks an apparent transmembrane sequence, which indicates that it may be a secreted, rather than a membrane-associated, molecule. We expressed a recombinant, histidine-tagged form of Dcr3 in mammalian cells; Dcr3 was secreted into the cell culture medium, and migrated on polyacrylamide gels as a protein of relative molecular mass 35,000 (data not shown). Dcr3 shares sequence identity in particular with OPG (31%) and TNFR2 (29%), and has relatively less homology with Fas (17%). All of the cysteines in the four CRDs of Dcr3 and OPG are conserved; however, the carboxy-terminal portion of Dcr3 is 101 residues shorter.

We analysed expression of Dcr3 mRNA in human tissues by northern blotting (Fig. 1b). We detected a predominant 1.2-kilobase transcript in fetal lung, brain, and liver, and in adult spleen, colon and lung. In addition, we observed relatively high Dcr3 mRNA expression in the human colon carcinoma cell line SW480.

To investigate potential ligand interactions of Dcr3, we generated a recombinant, Fc-tagged Dcr3 protein. We tested binding of Dcr3–Fc to human 293 cells transfected with individual TNF-family ligands, which are expressed as type 2 transmembrane proteins (these transmembrane proteins have their N termini in the cytosol). Dcr3–Fc showed a significant increase in binding to cells transfected with FasL<sup>4</sup> (Fig. 2a), but not to cells transfected with TNF<sup>5</sup>, Apo2L/TRAIL<sup>6,7</sup>, Apo3L/TWEAK<sup>8,9</sup>, or OPGL/TRACE/

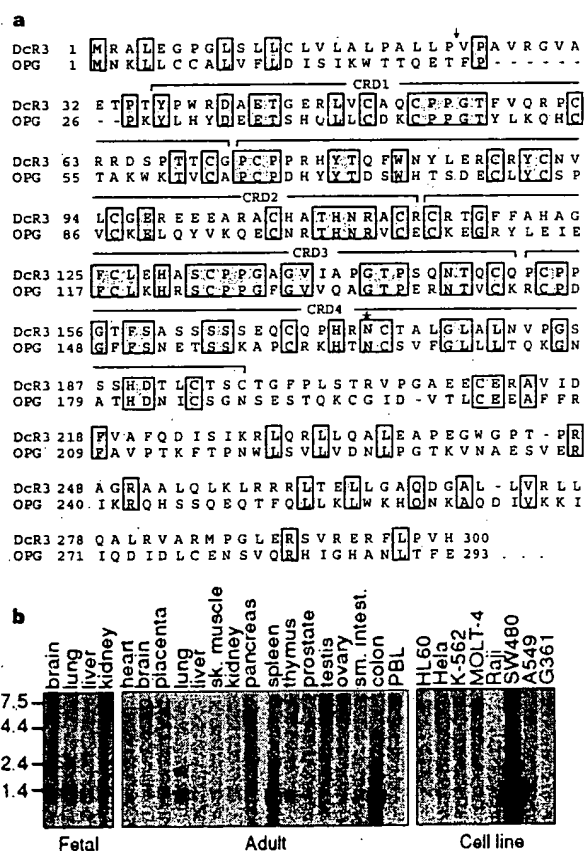
RANKL<sup>10-12</sup> (data not shown). DcR3-Fc immunoprecipitated shed FasL from FasL-transfected 293 cells (Fig. 2b) and purified soluble FasL (Fig. 2c), as did the Fc-tagged ectodomain of Fas but not TNFR1. Gel-filtration chromatography showed that DcR3-Fc and soluble FasL formed a stable complex (Fig. 2d). Equilibrium analysis indicated that DcR3-Fc and Fas-Fc bound to soluble FasL with a comparable affinity ( $K_d = 0.8 \pm 0.2$  and  $1.1 \pm 0.1$  nM, respectively; Fig. 2e), and that DcR3-Fc could block nearly all of the binding of soluble FasL to Fas-Fc (Fig. 2e, inset). Thus, DcR3 competes with Fas for binding to FasL.

To determine whether binding of DcR3 inhibits FasL activity, we tested the effect of DcR3-Fc on apoptosis induction by soluble FasL in Jurkat T leukaemia cells, which express Fas (Fig. 3a). DcR3-Fc and Fas-Fc blocked soluble-FasL-induced apoptosis in a similar dose-dependent manner, with half-maximal inhibition at  $\sim 0.1 \mu\text{g ml}^{-1}$ . Time-course analysis showed that the inhibition did not merely delay cell death, but rather persisted for at least 24 hours (Fig. 3b). We also tested the effect of DcR3-Fc on activation-induced cell death (AICD) of mature T lymphocytes, a FasL-dependent process<sup>1</sup>. Consistent with previous results<sup>13</sup>, activation of interleukin-2-stimulated CD4-positive T cells with anti-CD3 antibody increased the level of apoptosis twofold, and Fas-Fc blocked this effect substantially (Fig. 3c); DcR3-Fc blocked the

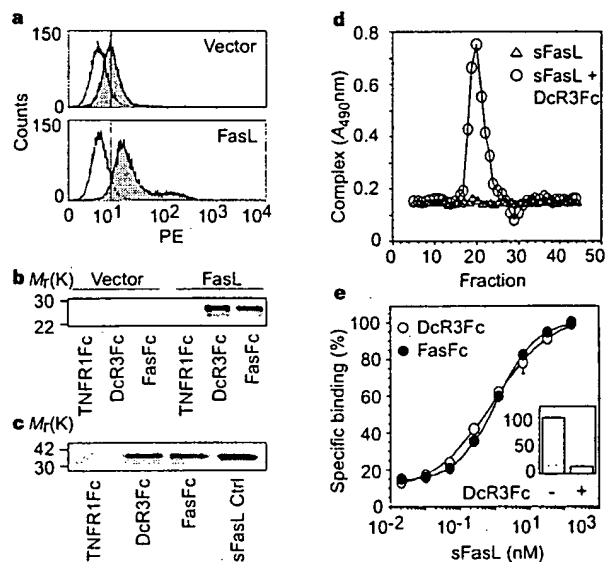
induction of apoptosis to a similar extent. Thus, DcR3 binding blocks apoptosis induction by FasL.

FasL-induced apoptosis is important in elimination of virus-infected cells and cancer cells by natural killer cells and cytotoxic T lymphocytes; an alternative mechanism involves perforin and granzymes<sup>1,14-16</sup>. Peripheral blood natural killer cells triggered marked cell death in Jurkat T leukaemia cells (Fig. 3d); DcR3-Fc and Fas-Fc each reduced killing of target cells from  $\sim 65\%$  to  $\sim 30\%$ , with half-maximal inhibition at  $\sim 1 \mu\text{g ml}^{-1}$ ; the residual killing was probably mediated by the perforin/granzyme pathway. Thus, DcR3 binding blocks FasL-dependent natural killer cell activity. Higher DcR3-Fc and Fas-Fc concentrations were required to block natural killer cell activity compared with those required to block soluble FasL activity, which is consistent with the greater potency of membrane-associated FasL compared with soluble FasL<sup>17</sup>.

Given the role of immune-cytotoxic cells in elimination of tumour cells and the fact that DcR3 can act as an inhibitor of FasL, we proposed that DcR3 expression might contribute to the ability of some tumours to escape immune-cytotoxic attack. As genomic amplification frequently contributes to tumorigenesis, we investigated whether the DcR3 gene is amplified in cancer. We analysed DcR3 gene-copy number by quantitative polymerase chain



**Figure 1** Primary structure and expression of human DcR3. **a**, Alignment of the amino-acid sequences of DcR3 and of osteoprotegerin (OPG); the C-terminal 101 residues of OPG are not shown. The putative signal cleavage site (arrow), the cysteine-rich domains (CRD 1-4), and the N-linked glycosylation site (asterisk) are shown. **b**, Expression of DcR3 mRNA. Northern hybridization analysis was done using the DcR3 cDNA as a probe and blots of poly(A)<sup>+</sup> RNA (Clontech) from human fetal and adult tissues or cancer cell lines. PBL, peripheral blood lymphocyte.



**Figure 2** Interaction of DcR3 with FasL. **a**, 293 cells were transfected with pRK5 vector (top) or with pRK5 encoding full-length FasL (bottom), incubated with DcR3-Fc (solid line, shaded area), TNFR1-Fc (dotted line) or buffer control (dashed line) (the dashed and dotted lines overlap), and analysed for binding by FACS. Statistical analysis showed a significant difference ( $P < 0.001$ ) between the binding of DcR3-Fc to cells transfected with FasL or pRK5. PE, phycoerythrin-labelled cells. **b**, 293 cells were transfected as in **a** and metabolically labelled, and cell supernatants were immunoprecipitated with Fc-tagged TNFR1, DcR3 or Fas. **c**, Purified soluble FasL (sFasL) was immunoprecipitated with TNFR1-Fc, DcR3-Fc or Fas-Fc and visualized by immunoblot with anti-FasL antibody. sFasL was loaded directly for comparison in the right-hand lane. **d**, Flag-tagged sFasL was incubated with DcR3-Fc or with buffer and resolved by gel filtration; column fractions were analysed in an assay that detects complexes containing DcR3-Fc and sFasL-Flag. **e**, Equilibrium binding of DcR3-Fc or Fas-Fc to sFasL-Flag. Inset, competition of DcR3-Fc with Fas-Fc for binding to sFasL-Flag.



reaction (PCR)<sup>18</sup> in genomic DNA from 35 primary lung and colon tumours, relative to pooled genomic DNA from peripheral blood leukocytes (PBLs) of 10 healthy donors. Eight of 18 lung tumours and 9 of 17 colon tumours showed DcR3 gene amplification, ranging from 2- to 18-fold (Fig. 4a, b). To confirm this result, we analysed the colon tumour DNAs with three more, independent sets of DcR3-based PCR primers and probes; we observed nearly the same amplification (data not shown).

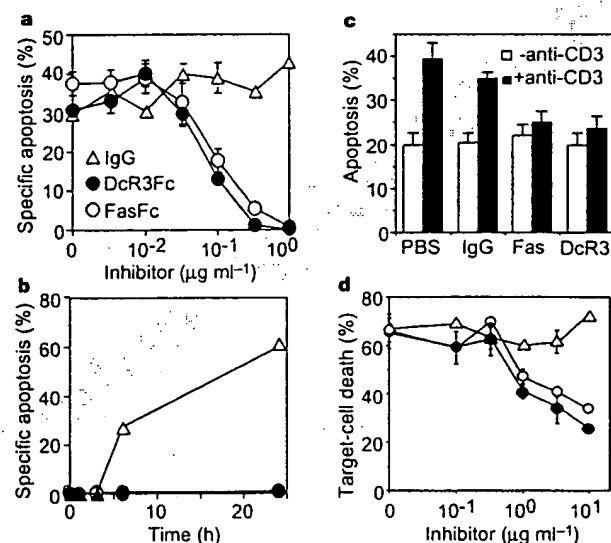
We then analysed DcR3 mRNA expression in primary tumour tissue sections by *in situ* hybridization. We detected DcR3 expression in 6 out of 15 lung tumours, 2 out of 2 colon tumours, 2 out of 5 breast tumours, and 1 out of 1 gastric tumour (data not shown). A section through a squamous-cell carcinoma of the lung is shown in Fig. 4c. DcR3 mRNA was localized to infiltrating malignant epithelium, but was essentially absent from adjacent stroma, indicating tumour-specific expression. Although the individual tumour specimens that we analysed for mRNA expression and gene amplification were different, the *in situ* hybridization results are consistent with the finding that the DcR3 gene is amplified frequently in tumours. SW480 colon carcinoma cells, which showed abundant DcR3 mRNA expression (Fig. 1b), also had marked DcR3 gene amplification, as shown by quantitative PCR (fourfold) and by Southern blot hybridization (fivefold) (data not shown).

If DcR3 amplification in cancer is functionally relevant, then DcR3 should be amplified more than neighbouring genomic regions that are not important for tumour survival. To test this,

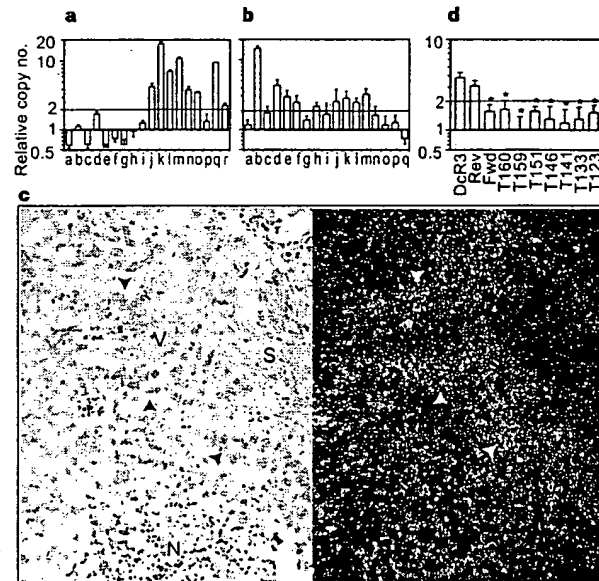
we mapped the human DcR3 gene by radiation-hybrid analysis; DcR3 showed linkage to marker AFM218x7 (T160), which maps to chromosome position 20q13. Next, we isolated from a bacterial artificial chromosome (BAC) library a human genomic clone that carries DcR3, and sequenced the ends of the clone's insert. We then determined, from the nine colon tumours that showed twofold or greater amplification of DcR3, the copy number of the DcR3-flanking sequences (reverse and forward) from the BAC, and of seven genomic markers that span chromosome 20 (Fig. 4d). The DcR3-linked reverse marker showed an average amplification of roughly threefold, slightly less than the approximately fourfold amplification of DcR3; the other markers showed little or no amplification. These data indicate that DcR3 may be at the 'epicentre' of a distal chromosome 20 region that is amplified in colon cancer, consistent with the possibility that DcR3 amplification promotes tumour survival.

Our results show that DcR3 binds specifically to FasL and inhibits FasL activity. We did not detect DcR3 binding to several other TNF-ligand-family members; however, this does not rule out the possibility that DcR3 interacts with other ligands, as do some other TNFR family members, including OPG<sup>2,19</sup>.

FasL is important in regulating the immune response; however, little is known about how FasL function is controlled. One mechanism involves the molecule cFLIP, which modulates apoptosis signalling downstream of Fas<sup>20</sup>. A second mechanism involves proteolytic shedding of FasL from the cell surface<sup>17</sup>. DcR3 competes with Fas for



**Figure 3** Inhibition of FasL activity by DcR3. **a**, Human Jurkat T leukaemia cells were incubated with Flag-tagged soluble FasL (sFasL; 5 ng ml<sup>-1</sup>) oligomerized with anti-Flag antibody (0.1 μg ml<sup>-1</sup>) in the presence of the proposed inhibitors DcR3-Fc, Fas-Fc or human IgG1 and assayed for apoptosis (mean ± s.e.m. of triplicates). **b**, Jurkat cells were incubated with sFasL-Flag plus anti-Flag antibody as in **a**, in presence of 1 μg ml<sup>-1</sup> DcR3-Fc (filled circles), Fas-Fc (open circles) or human IgG1 (triangles), and apoptosis was determined at the indicated time points. **c**, Peripheral blood T cells were stimulated with PHA and interleukin-2, followed by control (white bars) or anti-CD3 antibody (filled bars), together with phosphate-buffered saline (PBS), human IgG1, Fas-Fc, or DcR3-Fc (10 μg ml<sup>-1</sup>). After 16 h, apoptosis of CD4<sup>+</sup> cells was determined (mean ± s.e.m. of results from five donors). **d**, Peripheral blood natural killer cells were incubated with <sup>51</sup>Cr-labelled Jurkat cells in the presence of DcR3-Fc (filled circles), Fas-Fc (open circles) or human IgG1 (triangles), and target-cell death was determined by release of <sup>51</sup>Cr (mean ± s.d. for two donors, each in triplicate).



**Figure 4** Genomic amplification of DcR3 in tumours. **a**, Lung cancers, comprising eight adenocarcinomas (c, d, f, g, h, j, k, r), seven squamous-cell carcinomas (a, e, m, n, o, p, q), one non-small-cell carcinoma (b), one small-cell carcinoma (i), and one bronchial adenocarcinoma (l). The data are means ± s.d. of 2 experiments done in duplicate. **b**, Colon tumours, comprising 17 adenocarcinomas. Data are means ± s.e.m. of five experiments done in duplicate. **c**, *In situ* hybridization analysis of DcR3 mRNA expression in a squamous-cell carcinoma of the lung. A representative bright-field image (left) and the corresponding dark-field image (right) show DcR3 mRNA over infiltrating malignant epithelium (arrowheads). Adjacent non-malignant stroma (S), blood vessel (V) and necrotic tumour tissue (N) are also shown. **d**, Average amplification of DcR3 compared with amplification of neighbouring genomic regions (reverse and forward, Rev and Fwd), the DcR3-linked marker T160, and other chromosome-20 markers, in the nine colon tumours showing DcR3 amplification of twofold or more (b). Data are from two experiments done in duplicate. Asterisk indicates  $P < 0.01$  for a Student's *t*-test comparing each marker with DcR3.



FasL binding; hence, it may represent a third mechanism of extracellular regulation of FasL activity. A decoy receptor that modulates the function of the cytokine interleukin-1 has been described<sup>21</sup>. In addition, two decoy receptors that belong to the TNFR family, DcR1 and DcR2, regulate the FasL-related apoptosis-inducing molecule Apo2L<sup>22</sup>. Unlike DcR1 and DcR2, which are membrane-associated proteins, DcR3 is directly secreted into the extracellular space. One other secreted TNFR-family member is OPG<sup>3</sup>, which shares greater sequence homology with DcR3 (31%) than do DcR1 (17%) or DcR2 (19%); OPG functions as a third decoy for Apo2L<sup>19</sup>. Thus, DcR3 and OPG define a new subset of TNFR-family members that function as secreted decoys to modulate ligands that induce apoptosis. Pox viruses produce soluble TNFR homologues that neutralize specific TNF-family ligands, thereby modulating the antiviral immune response<sup>2</sup>. Our results indicate that a similar mechanism, namely, production of a soluble decoy receptor for FasL, may contribute to immune evasion by certain tumours. □

## Methods

**Isolation of DcR3 cDNA.** Several overlapping ESTs in GenBank (accession numbers AA025672, AA025673 and W67560) and in Lifeseq<sup>TM</sup> (Incyte Pharmaceuticals; accession numbers 1339238, 1533571, 1533650, 1542861, 1789372 and 2207027) showed similarity to members of the TNFR family. We screened human cDNA libraries by PCR with primers based on the region of EST consensus; fetal lung was positive for a product of the expected size. By hybridization to a PCR-generated probe based on the ESTs, one positive clone (DNA30942) was identified. When searching for potential alternatively spliced forms of DcR3 that might encode a transmembrane protein, we isolated 50 more clones; the coding regions of these clones were identical in size to that of the initial clone (data not shown).

**Fc-fusion proteins (immunoadhesins).** The entire DcR3 sequence, or the ectodomain of Fas or TNFR1, was fused to the hinge and Fc region of human IgG1, expressed in insect SF9 cells or in human 293 cells, and purified as described<sup>23</sup>.

**Fluorescence-activated cell sorting (FACS) analysis.** We transfected 293 cells using calcium phosphate or Effectene (Qiagen) with pRK5 vector or pRK5 encoding full-length human FasL<sup>4</sup> (2 µg), together with pRK5 encoding CrmA (2 µg) to prevent cell death. After 16 h, the cells were incubated with biotinylated DcR3-Fc or TNFR1-Fc and then with phycoerythrin-conjugated streptavidin (GibcoBRL), and were assayed by FACS. The data were analysed by Kolmogorov-Smirnov statistical analysis. There was some detectable staining of vector-transfected cells by DcR3-Fc; as these cells express little FasL (data not shown), it is possible that DcR3 recognized some other factor that is expressed constitutively on 293 cells.

**Immunoprecipitation.** Human 293 cells were transfected as above, and metabolically labelled with [<sup>35</sup>S]cysteine and [<sup>35</sup>S]methionine (0.5 mCi; Amersham). After 16 h of culture in the presence of z-VAD-fmk (10 µM), the medium was immunoprecipitated with DcR3-Fc, Fas-Fc or TNFR1-Fc (5 µg), followed by protein A-Sepharose (Repligen). The precipitates were resolved by SDS-PAGE and visualized on a phosphorimager (Fuji BAS2000). Alternatively, purified, Flag-tagged soluble FasL (1 µg) (Alexis) was incubated with each Fc-fusion protein (1 µg), precipitated with protein A-Sepharose, resolved by SDS-PAGE and visualized by immunoblotting with rabbit anti-FasL antibody (Oncogene Research).

**Analysis of complex formation.** Flag-tagged soluble FasL (25 µg) was incubated with buffer or with DcR3-Fc (40 µg) for 1.5 h at 24 °C. The reaction was loaded onto a Superdex 200 HR 10/30 column (Pharmacia) and developed with PBS; 0.6-ml fractions were collected. The presence of DcR3-Fc-FasL complex in each fraction was analysed by placing 100 µl aliquots into microtitre wells precoated with anti-human IgG (Boehringer) to capture DcR3-Fc, followed by detection with biotinylated anti-Flag antibody Bio M2 (Kodak) and streptavidin-horse radish peroxidase (Amersham). Calibration of the column indicated an apparent relative molecular mass of the complex of 420K (data not shown), which is consistent with a stoichiometry of two DcR3-Fc homodimers to two soluble FasL homotrimers.

**Equilibrium binding analysis.** Microtitre wells were coated with anti-human

IgG, blocked with 2% BSA in PBS. DcR3-Fc or Fas-Fc was added, followed by serially diluted Flag-tagged soluble FasL. Bound ligand was detected with anti-Flag antibody as above. In the competition assay, Fas-Fc was immobilized as above, and the wells were blocked with excess IgG1 before addition of Flag-tagged soluble FasL plus DcR3-Fc.

**T-cell AICD.** CD3<sup>+</sup> lymphocytes were isolated from peripheral blood of individual donors using anti-CD3 magnetic beads (Miltenyi Biotech), stimulated with phytohaemagglutinin (PHA; 2 µg ml<sup>-1</sup>) for 24 h, and cultured in the presence of interleukin-2 (100 U ml<sup>-1</sup>) for 5 days. The cells were plated in wells coated with anti-CD3 antibody (Pharmingen) and analysed for apoptosis 16 h later by FACS analysis of annexin-V-binding of CD4<sup>+</sup> cells<sup>24</sup>.

**Natural killer cell activity.** Natural killer cells were isolated from peripheral blood of individual donors using anti-CD56 magnetic beads (Miltenyi Biotech), and incubated for 16 h with <sup>51</sup>Cr-loaded Jurkat cells at an effector-to-target ratio of 1:1 in the presence of DcR3-Fc, Fas-Fc or human IgG1. Target-cell death was determined by release of <sup>51</sup>Cr in effector-target co-cultures relative to release of <sup>51</sup>Cr by detergent lysis of equal numbers of Jurkat cells.

**Gene-amplification analysis.** Surgical specimens were provided by J. Kern (lung tumours) and P. Quirke (colon tumours). Genomic DNA was extracted (Qiagen) and the concentration was determined using Hoechst dye 33258 intercalation fluorometry. Amplification was determined by quantitative PCR<sup>18</sup> using a TaqMan instrument (ABI). The method was validated by comparison of PCR and Southern hybridization data for the Myc and HER-2 oncogenes (data not shown). Gene-specific primers and fluorogenic probes were designed on the basis of the sequence of DcR3 or of nearby regions identified on a BAC carrying the human DcR3 gene; alternatively, primers and probes were based on Stanford Human Genome Center marker AFM218xe7 (T160), which is linked to DcR3 (likelihood score = 5.4), SHGC-36268 (T159), the nearest available marker which maps to ~500 kilobases from T160, and five extra markers that span chromosome 20. The DcR3-specific primer sequences were 5'-CTTCTTCGCGCAGCTG-3' and 5'-ATCAGCCGCGACACAG-3' and the fluorogenic probe sequence was 5'-(FAM-ACACGATGCGTGTCCAGCAG AAp-(TAMARA), where FAM is 5'-fluorescein phosphoramidite. Relative gene-copy numbers were derived using the formula 2<sup>(ΔCT)</sup>, where ΔCT is the difference in amplification cycles required to detect DcR3 in peripheral blood lymphocyte DNA compared to test DNA.

Received 24 September; accepted 6 November 1998.

1. Nagata, S. Apoptosis by death factor. *Cell* **88**, 355–365 (1997).
2. Smith, C. A., Farrah, T. & Goodwin, R. G. The TNF receptor superfamily of cellular and viral proteins: activation, costimulation, and death. *Cell* **76**, 959–962 (1994).
3. Simonet, W. S. *et al.* Osteoprotegerin: a novel secreted protein involved in the regulation of bone density. *Cell* **89**, 309–319 (1997).
4. Suda, T., Takahashi, T., Golstein, P. & Nagata, S. Molecular cloning and expression of Fas ligand, a novel member of the TNF family. *Cell* **75**, 1169–1178 (1993).
5. Pennica, D. *et al.* Human tumour necrosis factor: precursor structure, expression and homology to lymphotoxin. *Nature* **312**, 724–729 (1984).
6. Pitti, R. M. *et al.* Induction of apoptosis by Apo-2 ligand, a new member of the tumor necrosis factor receptor family. *J. Biol. Chem.* **271**, 12687–12690 (1996).
7. Wiley, S. R. *et al.* Identification and characterization of a new member of the TNF family that induces apoptosis. *Immunity* **3**, 673–682 (1995).
8. Marsters, S. A. *et al.* Identification of a ligand for the death-domain-containing receptor Apo3. *Curr. Biol.* **8**, 525–528 (1998).
9. Chicheportiche, Y. *et al.* TWEAK, a new secreted ligand in the TNF family that weakly induces apoptosis. *J. Biol. Chem.* **272**, 32401–32410 (1997).
10. Wong, B. R. *et al.* TRANCE is a novel ligand of the TNFR family that activates c-Jun-N-terminal kinase in T cells. *J. Biol. Chem.* **272**, 25190–25194 (1997).
11. Anderson, D. M. *et al.* A homolog of the TNF receptor and its ligand enhance T-cell growth and dendritic-cell function. *Nature* **390**, 175–179 (1997).
12. Lacey, D. L. *et al.* Osteoprotegerin ligand is a cytokine that regulates osteoclast differentiation and activation. *Cell* **93**, 165–176 (1998).
13. Dhein, J., Walczak, H., Baumler, C., Debatin, K. M. & Krammer, P. H. Autocrine T-cell suicide mediated by Apo1/Fas(CD95). *Nature* **373**, 438–441 (1995).
14. Arase, H., Arase, N. & Saito, T. Fas-mediated cytotoxicity by freshly isolated natural killer cells. *J. Exp. Med.* **181**, 1235–1238 (1995).
15. Medvedev, A. E. *et al.* Regulation of Fas and Fas ligand expression in NK cells by cytokines and the involvement of Fas ligand in NK/LAK cell-mediated cytotoxicity. *Cytokine* **9**, 394–404 (1997).
16. Moretta, A. Mechanisms in cell-mediated cytotoxicity. *Cell* **90**, 13–18 (1997).
17. Tanaka, M., Itai, T., Adachi, M. & Nagata, S. Downregulation of Fas ligand by shedding. *Nature Med.* **4**, 31–36 (1998).
18. Gelmini, S. *et al.* Quantitative PCR-based homogeneous assay with fluorogenic probes to measure c-erbB-2 oncogene amplification. *Clin. Chem.* **43**, 752–758 (1997).
19. Emery, J. G. *et al.* Osteoprotegerin is a receptor for the cytotoxic ligand TRAIL. *J. Biol. Chem.* **273**, 14363–14367 (1998).
20. Wallach, D. Placing death under control. *Nature* **388**, 123–125 (1997).
21. Colotta, F. *et al.* Interleukin-1 type II receptor: a decoy target for IL-1 that is regulated by IL-4. *Science* **261**, 472–475 (1993).

22. Ashkenazi, A. & Dixit, V. M. Death receptors: signaling and modulation. *Science* 281, 1305–1308 (1998).
23. Ashkenazi, A. & Chomow, S. M. Immunoadhesins as research tools and therapeutic agents. *Curr. Opin. Immunol.* 9, 195–200 (1997).
24. Marsters, S. *et al.* Activation of apoptosis by Apo-2 ligand is independent of FADD but blocked by CrmA. *Curr. Biol.* 6, 750–752 (1996).

Acknowledgements. We thank C. Clark, D. Pennica and V. Dixit for comments, and J. Kern and P. Quirke for tumour specimens.

Correspondence and requests for materials should be addressed to A.A. (e-mail: aa@gene.com). The GenBank accession number for the DcR3 cDNA sequence is AF104419.

## Crystal structure of the ATP-binding subunit of an ABC transporter

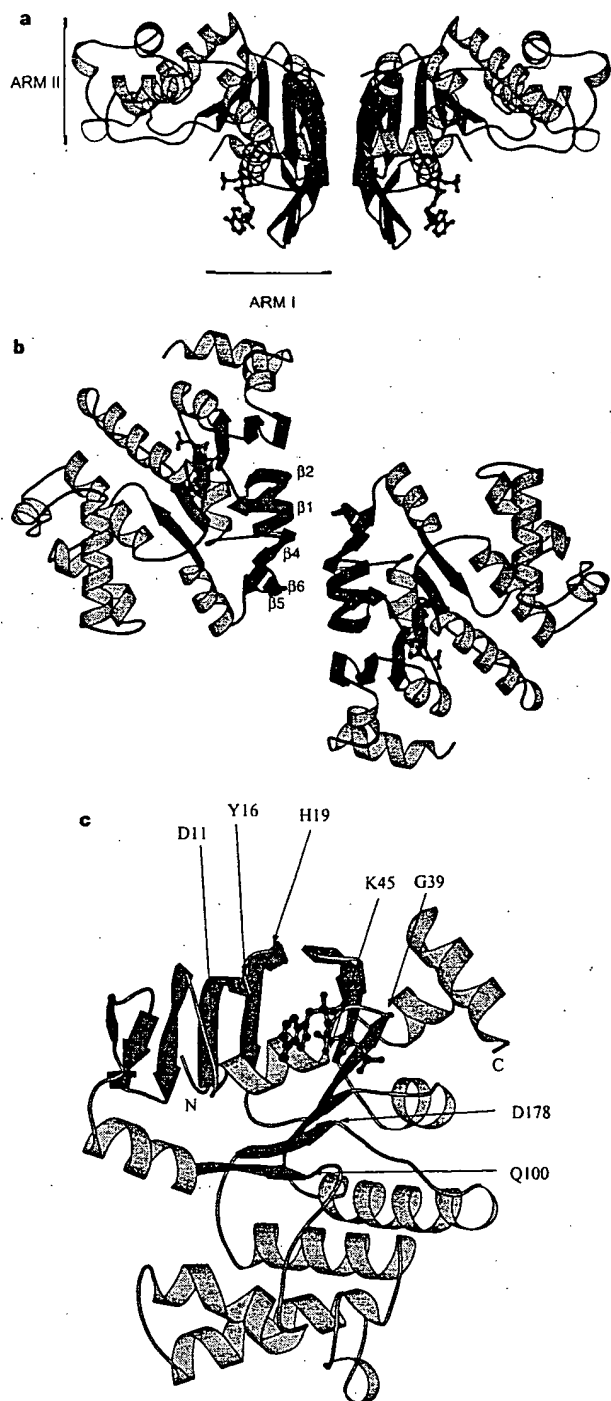
Li-Wei Hung\*, Iris Xiaoyan Wang†, Kishiko Nikaido†, Pei-Qi Liu†, Giovanna Ferro-Luzzi Ames† & Sung-Hou Kim\*‡

\* E. O. Lawrence Berkeley National Laboratory, † Department of Molecular and Cell Biology, and ‡ Department of Chemistry, University of California at Berkeley, Berkeley, California 94720, USA

ABC transporters (also known as traffic ATPases) form a large family of proteins responsible for the translocation of a variety of compounds across membranes of both prokaryotes and eukaryotes<sup>1</sup>. The recently completed *Escherichia coli* genome sequence revealed that the largest family of paralogous *E. coli* proteins is composed of ABC transporters<sup>2</sup>. Many eukaryotic proteins of medical significance belong to this family, such as the cystic fibrosis transmembrane conductance regulator (CFTR), the P-glycoprotein (or multidrug-resistance protein) and the heterodimeric transporter associated with antigen processing (Tap1–Tap2). Here we report the crystal structure at 1.5 Å resolution of HisP, the ATP-binding subunit of the histidine permease, which is an ABC transporter from *Salmonella typhimurium*. We correlate the details of this structure with the biochemical, genetic and biophysical properties of the wild-type and several mutant HisP proteins. The structure provides a basis for understanding properties of ABC transporters and of defective CFTR proteins.

ABC transporters contain four structural domains: two nucleotide-binding domains (NBDs), which are highly conserved throughout the family, and two transmembrane domains<sup>1</sup>. In prokaryotes these domains are often separate subunits which are assembled into a membrane-bound complex; in eukaryotes the domains are generally fused into a single polypeptide chain. The periplasmic histidine permease of *S. typhimurium* and *E. coli*<sup>3–8</sup> is a well-characterized ABC transporter that is a good model for this superfamily. It consists of a membrane-bound complex, HisQMP<sub>2</sub>, which comprises integral membrane subunits, HisQ and HisM, and two copies of HisP, the ATP-binding subunit. HisP, which has properties intermediate between those of integral and peripheral membrane proteins<sup>9</sup>, is accessible from both sides of the membrane, presumably by its interaction with HisQ and HisM<sup>6</sup>. The two HisP subunits form a dimer, as shown by their cooperativity in ATP hydrolysis<sup>5</sup>, the requirement for both subunits to be present for activity<sup>8</sup>, and the formation of a HisP dimer upon chemical cross-linking. Soluble HisP also forms a dimer<sup>3</sup>. HisP has been purified and characterized in an active soluble form<sup>3</sup> which can be reconstituted into a fully active membrane-bound complex<sup>8</sup>.

The overall shape of the crystal structure of the HisP monomer is that of an 'L' with two thick arms (arm I and arm II); the ATP-binding pocket is near the end of arm I (Fig. 1). A six-stranded  $\beta$ -sheet ( $\beta 3$  and  $\beta 8$ – $\beta 12$ ) spans both arms of the L, with a domain of  $\alpha$ - plus  $\beta$ -type structure ( $\beta 1$ ,  $\beta 2$ ,  $\beta 4$ – $\beta 7$ ,  $\alpha 1$  and  $\alpha 2$ ) on one side (within arm I) and a domain of mostly  $\alpha$ -helices ( $\alpha 3$ – $\alpha 9$ ) on the



**Figure 1** Crystal structure of HisP. **a**, View of the dimer along an axis perpendicular to its two-fold axis. The top and bottom of the dimer are suggested to face towards the periplasmic and cytoplasmic sides, respectively (see text). The thickness of arm II is about 25 Å, comparable to that of membrane.  $\alpha$ -Helices are shown in orange and  $\beta$ -sheets in green. **b**, View along the two-fold axis of the HisP dimer, showing the relative displacement of the monomers not apparent in **a**. The  $\beta$ -strands at the dimer interface are labelled. **c**, View of one monomer from the bottom of arm I, as shown in **a**, towards arm II, showing the ATP-binding pocket. **a**–**c**, The protein and the bound ATP are in 'ribbon' and 'ball-and-stick' representations, respectively. Key residues discussed in the text are indicated in **c**. These figures were prepared with MOLSCRIPT<sup>20</sup>. N, amino terminus; C, C terminus.

## NOVEL APPROACH TO QUANTITATIVE POLYMERASE CHAIN REACTION USING REAL-TIME DETECTION: APPLICATION TO THE DETECTION OF GENE AMPLIFICATION IN BREAST CANCER

Ivan BIÈCHE<sup>1,2</sup>, Martine OLIVI<sup>1</sup>, Marie-Hélène CHAMPÈME<sup>2</sup>, Dominique VIDAUD<sup>1</sup>, Rosette LIDÉREAU<sup>2</sup> and Michel VIDAUD<sup>1\*</sup>

<sup>1</sup>Laboratoire de Génétique Moléculaire, Faculté des Sciences Pharmaceutiques et Biologiques de Paris, Paris, France

<sup>2</sup>Laboratoire d'Oncogénétique, Centre René Huguenin, St-Cloud, France

Gene amplification is a common event in the progression of human cancers, and amplified oncogenes have been shown to have diagnostic, prognostic and therapeutic relevance. A kinetic quantitative polymerase-chain-reaction (PCR) method, based on fluorescent TaqMan methodology and a new instrument (ABI Prism 7700 Sequence Detection System) capable of measuring fluorescence in real-time, was used to quantify gene amplification in tumor DNA. Reactions are characterized by the point during cycling when PCR amplification is still in the exponential phase, rather than the amount of PCR product accumulated after a fixed number of cycles. None of the reaction components is limited during the exponential phase, meaning that values are highly reproducible in reactions starting with the same copy number. This greatly improves the precision of DNA quantification. Moreover, real-time PCR does not require post-PCR sample handling, thereby preventing potential PCR-product carry-over contamination; it possesses a wide dynamic range of quantification and results in much faster and higher sample throughput. The real-time PCR method, was used to develop and validate a simple and rapid assay for the detection and quantification of the 3 most frequently amplified genes (*myc*, *ccnd1* and *erbB2*) in breast tumors. Extra copies of *myc*, *ccnd1* and *erbB2* were observed in 10, 23 and 15%, respectively, of 108 breast-tumor DNA; the largest observed numbers of gene copies were 4.6, 18.6 and 15.1, respectively. These results correlated well with those of Southern blotting. The use of this new semi-automated technique will make molecular analysis of human cancers simpler and more reliable, and should find broad applications in clinical and research settings. *Int. J. Cancer* 78:661–666, 1998.

© 1998 Wiley-Liss, Inc.

Gene amplification plays an important role in the pathogenesis of various solid tumors, including breast cancer, probably because over-expression of the amplified target genes confers a selective advantage. The first technique used to detect genomic amplification was cytogenetic analysis. Amplification of several chromosome regions, visualized either as extrachromosomal double minutes (dmins) or as integrated homogeneously staining regions (HSRs), are among the main visible cytogenetic abnormalities in breast tumors. Other techniques such as comparative genomic hybridization (CGH) (Kallioniemi *et al.*, 1994) have also been used in broad searches for regions of increased DNA copy numbers in tumor cells, and have revealed some 20 amplified chromosome regions in breast tumors. Positional cloning efforts are underway to identify the critical gene(s) in each amplified region. To date, genes known to be amplified frequently in breast cancers include *myc* (8q24), *ccnd1* (11q13), and *erbB2* (17q12-q21) (for review, see Bièche and Lidereau, 1995).

Amplification of the *myc*, *ccnd1*, and *erbB2* proto-oncogenes should have clinical relevance in breast cancer, since independent studies have shown that these alterations can be used to identify sub-populations with a worse prognosis (Berns *et al.*, 1992; Schuuring *et al.*, 1992; Slamon *et al.*, 1987). Muss *et al.* (1994) suggested that these gene alterations may also be useful for the prediction and assessment of the efficacy of adjuvant chemotherapy and hormone therapy.

However, published results diverge both in terms of the frequency of these alterations and their clinical value. For instance, over 500 studies in 10 years have failed to resolve the controversy

surrounding the link suggested by Slamon *et al.* (1987) between *erbB2* amplification and disease progression. These discrepancies are partly due to the clinical, histological and ethnic heterogeneity of breast cancer, but technical considerations are also probably involved.

Specific genes (DNA) were initially quantified in tumor cells by means of blotting procedures such as Southern and slot blotting. These batch techniques require large amounts of DNA (5–10 µg/reaction) to yield reliable quantitative results. Furthermore, meticulous care is required at all stages of the procedures to generate blots of sufficient quality for reliable dosage analysis. Recently, PCR has proven to be a powerful tool for quantitative DNA analysis, especially with minimal starting quantities of tumor samples (small, early-stage tumors and formalin-fixed, paraffin-embedded tissues).

Quantitative PCR can be performed by evaluating the amount of product either after a given number of cycles (end-point quantitative PCR) or after a varying number of cycles during the exponential phase (kinetic quantitative PCR). In the first case, an internal standard distinct from the target molecule is required to ascertain PCR efficiency. The method is relatively easy but implies generating, quantifying and storing an internal standard for each gene studied. Nevertheless, it is the most frequently applied method to date.

One of the major advantages of the kinetic method is its rapidity in quantifying a new gene, since no internal standard is required (an external standard curve is sufficient). Moreover, the kinetic method has a wide dynamic range (at least 5 orders of magnitude), giving an accurate value for samples differing in their copy number. Unfortunately, the method is cumbersome and has therefore been rarely used. It involves aliquot sampling of each assay mix at regular intervals and quantifying, for each aliquot, the amplification product. Interest in the kinetic method has been stimulated by a novel approach using fluorescent TaqMan methodology and a new instrument (ABI Prism 7700 Sequence Detection System) capable of measuring fluorescence in real time (Gibson *et al.*, 1996; Heid *et al.*, 1996). The TaqMan reaction is based on the 5' nuclease assay first described by Holland *et al.* (1991). The latter uses the 5' nuclease activity of Taq polymerase to cleave a specific fluorogenic oligonucleotide probe during the extension phase of PCR. The approach uses dual-labeled fluorogenic hybridization probes (Lee *et al.*, 1993). One fluorescent dye, co-valently linked to the 5' end of the oligonucleotide, serves as a reporter [FAM (*i.e.*, 6-carboxy-fluorescein)] and its emission spectrum is quenched by a second fluorescent dye, TAMRA (*i.e.*, 6-carboxy-tetramethyl-rhodamine) attached to the 3' end. During the extension phase of the PCR

Grant sponsors: Association Pour la Recherche sur le Cancer and Ministère de l'Enseignement Supérieur et de la Recherche.

\*Correspondence to: Laboratoire de Génétique Moléculaire, Faculté des Sciences Pharmaceutiques et Biologiques de Paris, 4 Avenue de l'Observatoire, F-75006 Paris, France. Fax: (33)1-4407-1754. E-mail: mvidaoud@teaser.fr

Received 2 May 1998; Revised 30 June 1998

cycle, the fluorescent hybridization probe is hydrolyzed by the 5'-3' nucleolytic activity of DNA polymerase. Nuclease degradation of the probe releases the quenching of FAM fluorescence emission, resulting in an increase in peak fluorescence emission. The fluorescence signal is normalized by dividing the emission intensity of the reporter dye (FAM) by the emission intensity of a reference dye (i.e., ROX, 6-carboxy-X-rhodamine) included in TaqMan buffer, to obtain a ratio defined as the  $R_n$  (normalized reporter) for a given reaction tube. The use of a sequence detector enables the fluorescence spectra of all 96 wells of the thermal cycler to be measured continuously during PCR amplification.

The real-time PCR method offers several advantages over other current quantitative PCR methods (Celi *et al.*, 1994): (i) the probe-based homogeneous assay provides a real-time method for detecting only specific amplification products, since specific hybridization of both the primers and the probe is necessary to generate a signal; (ii) the  $C_t$  (threshold cycle) value used for quantification is measured when PCR amplification is still in the log phase of PCR product accumulation. This is the main reason why  $C_t$  is a more reliable measure of the starting copy number than are end-point measurements, in which a slight difference in a limiting component can have a drastic effect on the amount of product; (iii) use of  $C_t$  values gives a wider dynamic range (at least 5 orders of magnitude), reducing the need for serial dilution; (iv) The real-time PCR method is run in a closed-tube system and requires no post-PCR sample handling, thus avoiding potential contamination; (v) the system is highly automated, since the instrument continuously measures fluorescence in all 96 wells of the thermal cycler during PCR amplification and the corresponding software processes, and analyzes the fluorescence data; (vi) the assay is rapid, as results are available just one minute after thermal cycling is complete; (vii) the sample throughput of the method is high, since 96 reactions can be analyzed in 2 hr.

Here, we applied this semi-automated procedure to determine the copy numbers of the 3 most frequently amplified genes in breast tumors (*myc*, *ccnd1* and *erbB2*), as well as 2 genes (*alb* and *app*) located in a chromosome region in which no genetic changes have been observed in breast tumors. The results for 108 breast tumors were compared with previous Southern-blot data for the same samples.

#### MATERIAL AND METHODS

##### Tumor and blood samples

Samples were obtained from 108 primary breast tumors removed surgically from patients at the Centre René Huguénin; none of the patients had undergone radiotherapy or chemotherapy. Immediately after surgery, the tumor samples were placed in liquid nitrogen until extraction of high-molecular-weight DNA. Patients were included in this study if the tumor sample used for DNA preparation contained more than 60% of tumor cells (histological analysis). A blood sample was also taken from 18 of the same patients.

DNA was extracted from tumor tissue and blood leukocytes according to standard methods.

##### Real-time PCR

**Theoretical basis.** Reactions are characterized by the point during cycling when amplification of the PCR product is first detected, rather than by the amount of PCR product accumulated after a fixed number of cycles. The higher the starting copy number of the genomic DNA target, the earlier a significant increase in fluorescence is observed. The parameter  $C_t$  (threshold cycle) is defined as the fractional cycle number at which the fluorescence generated by cleavage of the probe passes a fixed threshold above baseline. The target gene copy number in unknown samples is quantified by measuring  $C_t$  and by using a standard curve to determine the starting copy number. The precise amount of genomic DNA (based on optical density) and its quality (i.e., lack

of extensive degradation) are both difficult to assess. We therefore also quantified a control gene (*alb*) mapping to chromosome region 4q11-q13, in which no genetic alterations have been found in breast-tumor DNA by means of CGH (Kallioniemi *et al.*, 1994).

Thus, the ratio of the copy number of the target gene to the copy number of the *alb* gene normalizes the amount and quality of genomic DNA. The ratio defining the level of amplification is termed "N", and is determined as follows:

$$N = \frac{\text{copy number of target gene (app, myc, ccnd1, erbB2)}}{\text{copy number of reference gene (alb)}}$$

**Primers, probes, reference human genomic DNA and PCR consumables.** Primers and probes were chosen with the assistance of the computer programs Oligo 4.0 (National Biosciences, Plymouth, MN), EuGene (Daniben Systems, Cincinnati, OH) and Primer Express (Perkin-Elmer Applied Biosystems, Foster City, CA).

Primers were purchased from DNAgency (Malvern, PA) and probes from Perkin-Elmer Applied Biosystems.

Nucleotide sequences for the oligonucleotide hybridization probes and primers are available on request.

The TaqMan PCR Core reagent kit, MicroAmp optical tubes, and MicroAmp caps were from Perkin-Elmer Applied Biosystems.

**Standard-curve construction.** The kinetic method requires a standard curve. The latter was constructed with serial dilutions of specific PCR products, according to Piatak *et al.* (1993). In practice, each specific PCR product was obtained by amplifying 20 ng of a standard human genomic DNA (Boehringer, Mannheim, Germany) with the same primer pairs as those used later for real-time quantitative PCR. The 5 PCR products were purified using MicroSpin S-400 HR columns (Pharmacia, Uppsala, Sweden) electrophoresed through an acrylamide gel and stained with ethidium bromide to check their quality. The PCR products were then quantified spectrophotometrically and pooled, and serially diluted 10-fold in mouse genomic DNA (Clontech, Palo Alto, CA) at a constant concentration of 2 ng/ $\mu$ l. The standard curve used for real-time quantitative PCR was based on serial dilutions of the pool of PCR products ranging from  $10^{-7}$  ( $10^5$  copies of each gene) to  $10^{-10}$  ( $10^2$  copies). This series of diluted PCR products was aliquoted and stored at  $-80^\circ\text{C}$  until use.

The standard curve was validated by analyzing 2 known quantities of calibrator human genomic DNA (20 ng and 50 ng).

**PCR amplification.** Amplification mixes (50  $\mu$ l) contained the sample DNA (around 20 ng, around 6600 copies of disomic genes),  $10\times$  TaqMan buffer (5  $\mu$ l), 200  $\mu$ M dATP, dCTP, dGTP, and 400  $\mu$ M dUTP, 5 mM  $\text{MgCl}_2$ , 1.25 units of AmpliTaq Gold, 0.5 units of AmpErase uracil N-glycosylase (UNG), 200 nM each primer and 100 nM probe. The thermal cycling conditions comprised 2 min at  $50^\circ\text{C}$  and 10 min at  $95^\circ\text{C}$ . Thermal cycling consisted of 40 cycles at  $95^\circ\text{C}$  for 15 s and  $65^\circ\text{C}$  for 1 min. Each assay included: a standard curve (from  $10^5$  to  $10^2$  copies) in duplicate, a no-template control, 20 ng and 50 ng of calibrator human genomic DNA (Boehringer) in triplicate, and about 20 ng of unknown genomic DNA in triplicate (26 samples can thus be analyzed on a 96-well microplate). All samples with a coefficient of variation (CV) higher than 10% were retested.

All reactions were performed in the ABI Prism 7700 Sequence Detection System (Perkin-Elmer Applied Biosystems), which detects the signal from the fluorogenic probe during PCR.

**Equipment for real-time detection.** The 7700 system has a built-in thermal cycler and a laser directed via fiber optical cables to each of the 96 sample wells. A charge-coupled-device (CDD) camera collects the emission from each sample and the data are analyzed automatically. The software accompanying the 7700 system calculates  $C_t$  and determines the starting copy number in the samples.

**Determination of gene amplification.** Gene amplification was calculated as described above. Only samples with an N value higher than 2 were considered to be amplified.

### RESULTS

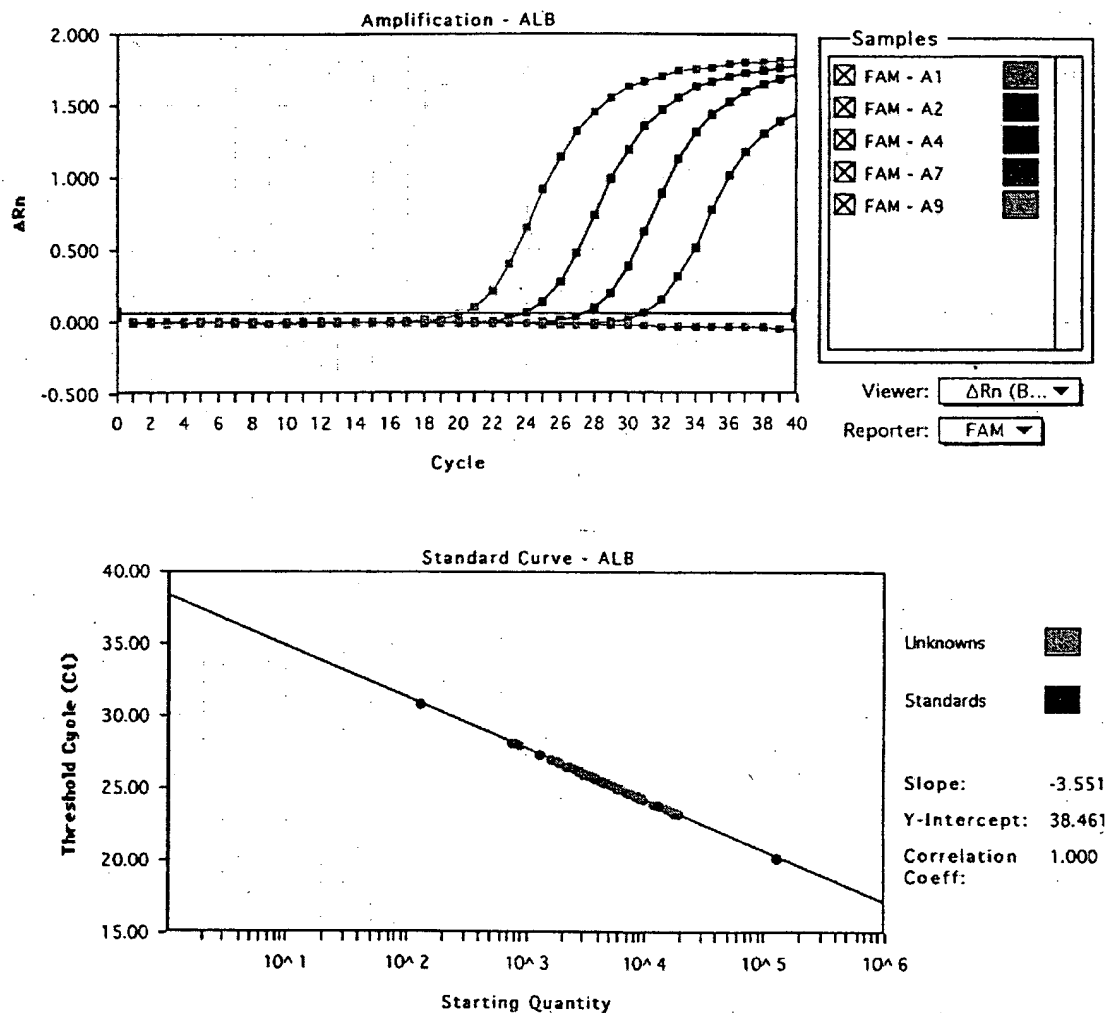
To validate the method, real-time PCR was performed on genomic DNA extracted from 108 primary breast tumors, and 18 normal leukocyte DNA samples from some of the same patients. The target genes were the *myc*, *ccnd1* and *erbB2* proto-oncogenes, and the  $\beta$ -amyloid precursor protein gene (*app*), which maps to a chromosome region (21q21.2) in which no genetic alterations have been found in breast tumors (Kallioniemi *et al.*, 1994). The reference disomic gene was the albumin gene (*alb*, chromosome 4q11-q13).

### Validation of the standard curve and dynamic range of real-time PCR

The standard curve was constructed from PCR products serially diluted in genomic mouse DNA at a constant concentration of 2 ng/ $\mu$ l. It should be noted that the 5 primer pairs chosen to analyze the 5 target genes do not amplify genomic mouse DNA (data not shown). Figure 1 shows the real-time PCR standard curve for the *alb* gene. The dynamic range was wide (at least 4 orders of magnitude), with samples containing as few as  $10^2$  copies or as many as  $10^5$  copies.

### Copy-number ratio of the 2 reference genes (*app* and *alb*)

The *app* to *alb* copy-number ratio was determined in 18 normal leukocyte DNA samples and all 108 primary breast-tumor DNA



**FIGURE 1** – Albumin (*alb*) gene dosage by real-time PCR. Top: Amplification plots for reactions with starting *alb* gene copy number ranging from  $10^5$  (A9),  $10^4$  (A7),  $10^3$  (A4) to  $10^2$  (A2) and a no-template control (A1). Cycle number is plotted vs. change in normalized reporter signal ( $\Delta Rn$ ). For each reaction tube, the fluorescence signal of the reporter dye (FAM) is divided by the fluorescence signal of the passive reference dye (ROX), to obtain a ratio defined as the normalized reporter signal (Rn).  $\Delta Rn$  represents the normalized reporter signal (Rn) minus the baseline signal established in the first 15 PCR cycles.  $\Delta Rn$  increases during PCR as *alb* PCR product copy number increases until the reaction reaches a plateau.  $C_t$  (threshold cycle) represents the fractional cycle number at which a significant increase in Rn above a baseline signal (horizontal black line) can first be detected. Two replicate plots were performed for each standard sample, but the data for only one are shown here. Bottom: Standard curve plotting log starting copy number vs.  $C_t$  (threshold cycle). The black dots represent the data for standard samples plotted in duplicate and the red dots the data for unknown genomic DNA samples plotted in triplicate. The standard curve shows 4 orders of linear dynamic range.

samples. We selected these 2 genes because they are located in 2 chromosome regions (*app*, 21q21.2; *alb*, 4q11-q13) in which no obvious genetic changes (including gains or losses) have been observed in breast cancers (Kallioniemi *et al.*, 1994). The ratio for the 18 normal leukocyte DNA samples fell between 0.7 and 1.3 (mean  $1.02 \pm 0.21$ ), and was similar for the 108 primary breast-tumor DNA samples (0.6 to 1.6, mean  $1.06 \pm 0.25$ ), confirming that *alb* and *app* are appropriate reference disomic genes for breast-tumor DNA. The low range of the ratios also confirmed that the nucleotide sequences chosen for the primers and probes were not polymorphic, as mismatches of their primers or probes with the subject's DNA would have resulted in differential amplification.

#### *myc*, *ccnd1* and *erbB2* gene dose in normal leukocyte DNA

To determine the cut-off point for gene amplification in breast-cancer tissue, 18 normal leukocyte DNA samples were tested for the gene dose (N), calculated as described in "Material and Methods". The N value of these samples ranged from 0.5 to 1.3 (mean  $0.84 \pm 0.22$ ) for *myc*, 0.7 to 1.6 (mean  $1.06 \pm 0.23$ ) for *ccnd1* and 0.6 to 1.3 (mean  $0.91 \pm 0.19$ ) for *erbB2*. Since N values for *myc*, *ccnd1* and *erbB2* in normal leukocyte DNA consistently fell between 0.5 and 1.6, values of 2 or more were considered to represent gene amplification in tumor DNA.

#### *myc*, *ccnd1* and *erbB2* gene dose in breast-tumor DNA

*myc*, *ccnd1* and *erbB2* gene copy numbers in the 108 primary breast tumors are reported in Table I. Extra copies of *ccnd1* were more frequent (23%, 25/108) than extra copies of *erbB2* (15%, 16/108) and *myc* (10%, 11/108), and ranged from 2 to 18.6 for *ccnd1*, 2 to 15.1 for *erbB2*, and only 2 to 4.6 for the *myc* gene. Figure 2 and Table II represent tumors in which the *ccnd1* gene was amplified 16-fold (T145), 6-fold (T133) and non-amplified (T118). The 3 genes were never found to be co-amplified in the same tumor. *erbB2* and *ccnd1* were co-amplified in only 3 cases, *myc* and *ccnd1* in 2 cases and *myc* and *erbB2* in 1 case. This favors the hypothesis that gene amplifications are independent events in breast cancer. Interestingly, 5 tumors showed a decrease of at least 50% in the *erbB2* copy number ( $N < 0.5$ ), suggesting that they bore deletions of the 17q21 region (the site of *erbB2*). No such decrease in copy number was observed with the other 2 proto-oncogenes.

#### Comparison of gene dose determined by real-time quantitative PCR and Southern-blot analysis

Southern-blot analysis of *myc*, *ccnd1* and *erbB2* amplifications had previously been done on the same 108 primary breast tumors. A perfect correlation between the results of real-time PCR and Southern blot was obtained for tumors with high copy numbers ( $N \geq 5$ ). However, there were cases (1 *myc*, 6 *ccnd1* and 4 *erbB2*) in which real-time PCR showed gene amplification whereas Southern-blot did not, but these were mainly cases with low extra copy numbers (N from 2 to 2.9).

### DISCUSSION

The clinical applications of gene amplification assays are currently limited, but would certainly increase if a simple, standardized and rapid method were perfected. Gene amplification status has been studied mainly by means of Southern blotting, but this method is not sensitive enough to detect low-level gene amplification nor accurate enough to quantify the full range of amplification values. Southern blotting is also time-consuming, uses radioactive

reagents and requires relatively large amounts of high-quality genomic DNA, which means it cannot be used routinely in many laboratories. An amplification step is therefore required to determine the copy number of a given target gene from minimal quantities of tumor DNA (small early-stage tumors, cytopuncture specimens or formalin-fixed, paraffin-embedded tissues).

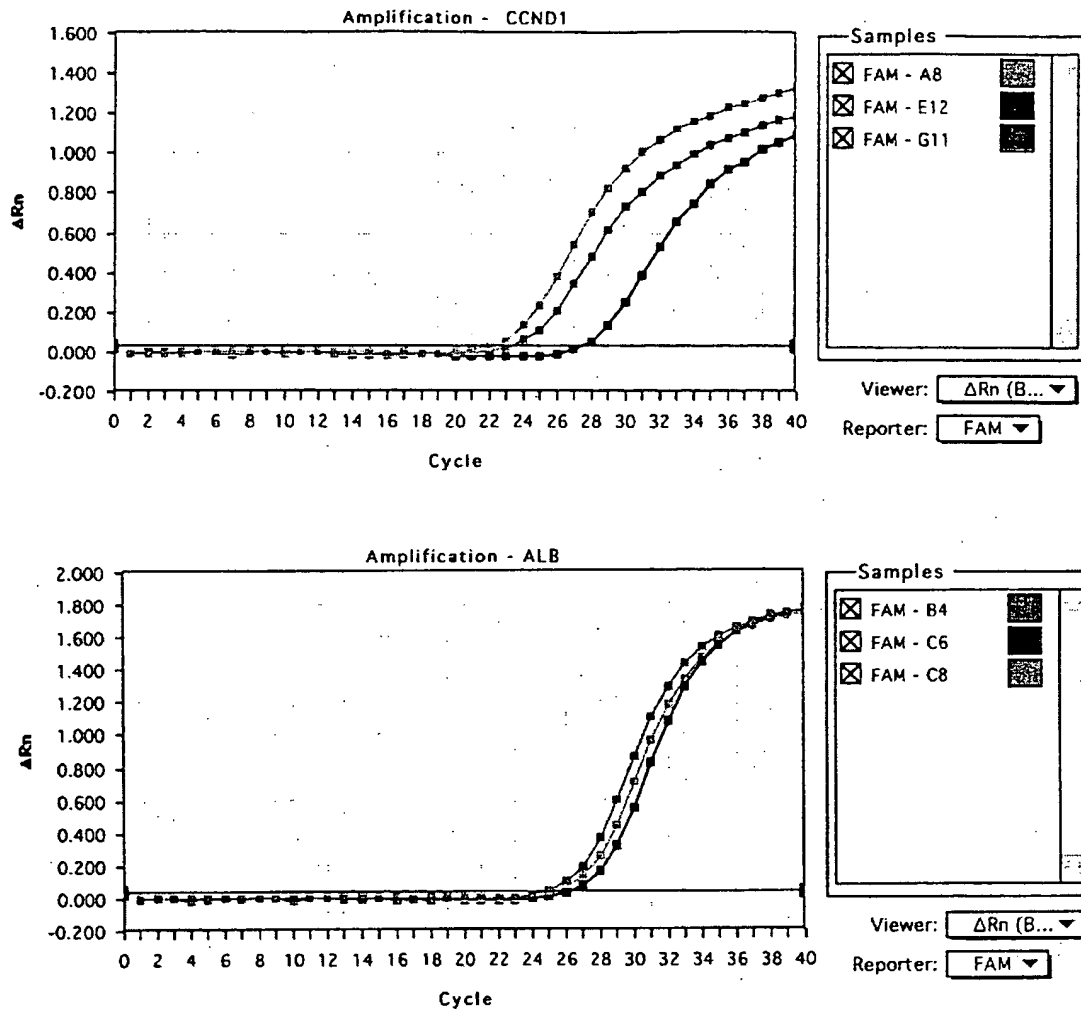
In this study, we validated a PCR method developed for the quantification of gene over-representation in tumors. The method, based on real-time analysis of PCR amplification, has several advantages over other PCR-based quantitative assays such as competitive quantitative PCR (Celi *et al.*, 1994). First, the real-time PCR method is performed in a closed-tube system, avoiding the risk of contamination by amplified products. Re-amplification of carryover PCR products in subsequent experiments can also be prevented by using the enzyme uracil N-glycosylase (UNG) (Longo *et al.*, 1990). The second advantage is the simplicity and rapidity of sample analysis, since no post-PCR manipulations are required. Our results show that the automated method is reliable. We found it possible to determine, in triplicate, the number of copies of a target gene in more than 100 tumors per day. Third, the system has a linear dynamic range of at least 4 orders of magnitude, meaning that samples do not have to contain equal starting amounts of DNA. This technique should therefore be suitable for analyzing formalin-fixed, paraffin-embedded tissues. Fourth, and above all, real-time PCR makes DNA quantification much more precise and reproducible, since it is based on  $C_t$  values rather than end-point measurement of the amount of accumulated PCR product. Indeed, the ABI Prism 7700 Sequence Detection System enables  $C_t$  to be calculated when PCR amplification is still in the exponential phase and when none of the reaction components is rate-limiting. The within-run CV of the  $C_t$  value for calibrator human DNA (5 replicates) was always below 5%, and the between-assay precision in 5 different runs was always below 10% (data not shown). In addition, the use of a standard curve is not absolutely necessary, since the copy number can be determined simply by comparing the  $C_t$  ratio of the target gene with that of reference genes. The results obtained by the 2 methods (with and without a standard curve) are similar in our experiments (data not shown). Moreover, unlike competitive quantitative PCR, real-time PCR does not require an internal control (the design and storage of internal controls and the validation of their amplification efficiency is laborious).

The only potential disadvantage of real-time PCR, like all other PCR-based methods and solid-matrix blotting techniques (Southern blots and dot blots) is that it cannot avoid dilution artifacts inherent in the extraction of DNA from tumor cells contained in heterogeneous tissue specimens. Only FISH and immunohistochemistry can measure alterations on a cell-by-cell basis (Pauletti *et al.*, 1996; Slamon *et al.*, 1989). However, FISH requires expensive equipment and trained personnel and is also time-consuming. Moreover, FISH does not assess gene expression and therefore cannot detect cases in which the gene product is over-expressed in the absence of gene amplification, which will be possible in the future by real-time quantitative RT-PCR. Immunohistochemistry is subject to considerable variations in the hands of different teams, owing to alterations of target proteins during the procedure, the different primary antibodies and fixation methods used and the criteria used to define positive staining.

The results of this study are in agreement with those reported in the literature. (i) Chromosome regions 4q11-q13 and 21q21.2 (which bear *alb* and *app*, respectively) showed no genetic alterations in the breast-cancer samples studied here, in keeping with the results of CGH (Kallioniemi *et al.*, 1994). (ii) We found that amplifications of these 3 oncogenes were independent events, as reported by other teams (Berns *et al.*, 1992; Borg *et al.*, 1992). (iii) The frequency and degree of *myc* amplification in our breast tumor DNA series were lower than those of *ccnd1* and *erbB2* amplification, confirming the findings of Borg *et al.* (1992) and Courjal *et al.* (1997). (iv) The maxima of *ccnd1* and *erbB2* over-representation were 18-fold and 15-fold, also in keeping with earlier results (about

TABLE I - DISTRIBUTION OF AMPLIFICATION LEVEL (N) FOR *myc*, *ccnd1* AND *erbB2* GENES IN 108 HUMAN BREAST TUMORS

Gene	Amplification level (N)			
	<0.5	0.5-1.9	2-4.9	$\geq 5$
<i>myc</i>	0	97 (89.8%)	11 (10.2%)	0
<i>ccnd1</i>	0	83 (76.9%)	17 (15.7%)	8 (7.4%)
<i>erbB2</i>	5 (4.6%)	87 (80.6%)	8 (7.4%)	8 (7.4%)



Tumor	CCND1		ALB	
	$C_t$	Copy number	$C_t$	Copy number
■ T118	27.3	4605	26.5	4365
■ T133	23.2	61659	25.2	10092
■ T145	22.1	125892	25.6	7762

FIGURE 2 - *ccnd1* and *alb* gene dosage by real-time PCR in 3 breast tumor samples: T118 (E12, C6, black squares), T133 (G11, B4, red squares) and T145 (A8, C8, blue squares). Given the  $C_t$  of each sample, the initial copy number is inferred from the standard curve obtained during the same experiment. Triplicate plots were performed for each tumor sample, but the data for only one are shown here. The results are shown in Table II.

30-fold maximum) (Berns *et al.*, 1992; Borg *et al.*, 1992; Courjal *et al.*, 1997). (v) The *erbB2* copy numbers obtained with real-time PCR were in good agreement with data obtained with other quantitative PCR-based assays in terms of the frequency and degree of amplification (An *et al.*, 1995; Deng *et al.*, 1996; Valeron

*et al.*, 1996). Our results also correlate well with those recently published by Gelmini *et al.* (1997), who used the TaqMan system to measure *erbB2* amplification in a small series of breast tumors ( $n = 25$ ), but with an instrument (LS-50B luminescence spectrometer, Perkin-Elmer Applied Biosystems) which only allows end-

TABLE II - EXAMPLES OF *ccnd1* GENE DOSAGE RESULTS FROM 3 BREAST TUMORS<sup>1</sup>

Tumor	<i>ccnd1</i>			<i>alb</i>			<i>Nccnd1/alb</i>
	Copy number	Mean	SD	Copy number	Mean	SD	
T118	4525	4603	77	4223	4325	89	1.06
	4605			4365			
	4678			4387			
T133	59821	61100	1111	9787	10092	375	6.03
	61659			10092			
	61821			10533			
T145	128563	125392	3448	7321	7672	316	16.34
	125892			7762			
	121722			7933			

<sup>1</sup>For each sample, 3 replicate experiments were performed and the mean and the standard deviation (SD) was determined. The level of *ccnd1* gene amplification (*Nccnd1/alb*) is determined by dividing the average *ccnd1* copy number value by the average *alb* copy number value.

point measurement of fluorescence intensity. Here we report *myc* and *ccnd1* gene dosage in breast cancer by means of quantitative PCR. (vi) We found a high degree of concordance between real-time quantitative PCR and Southern blot analysis in terms of gene amplification, especially for samples with high copy numbers ( $\geq 5$ -fold). The slightly higher frequency of gene amplification (especially *ccnd1* and *erbB2*) observed by means of real-time quantitative PCR as compared with Southern-blot analysis may be explained by the higher sensitivity of the former method. However, we cannot rule out the possibility that some tumors with a few extra

gene copies observed in real-time PCR had additional copies of an arm or a whole chromosome (trisomy, tetrasomy or polysomy) rather than true gene amplification. These 2 types of genetic alteration (polysomy and gene amplification) could be easily distinguished in the future by using an additional probe located on the same chromosome arm, but some distance from the target gene. It is noteworthy that high gene copy numbers have the greatest prognostic significance in breast carcinoma (Borg *et al.*, 1992; Slamon *et al.*, 1987).

Finally, this technique can be applied to the detection of gene deletion as well as gene amplification. Indeed, we found a decreased copy number of *erbB2* (but not of the other 2 proto-oncogenes) in several tumors; *erbB2* is located in a chromosome region (17q21) reported to contain both deletions and amplifications in breast cancer (Bièche and Lidereau, 1995).

In conclusion, gene amplification in various cancers can be used as a marker of pre-neoplasia, also for early diagnosis of cancer, staging, prognostication and choice of treatment. Southern blotting is not sufficiently sensitive, and FISH is lengthy and complex. Real-time quantitative PCR overcomes both these limitations, and is a sensitive and accurate method of analyzing large numbers of samples in a short time. It should find a place in routine clinical gene dosage.

#### ACKNOWLEDGEMENTS

RL is a research director at the Institut National de la Santé et de la Recherche Médicale (INSERM). We thank the staff of the Centre René Huguenin for assistance in specimen collection and patient care.

#### REFERENCES

- AN, H.X., NIEDERACHER, D., BECKMANN, M.W., GÖHRING, U.J., SCHARL, A., PICARD, F., VAN ROEYEN, C., SCHNÜRCH, H.G. and BENDER, H.G., *erbB2* gene amplification detected by fluorescent differential polymerase chain reaction in paraffin-embedded breast carcinoma tissues. *Int. J. Cancer (Pred. Oncol.)*, **64**, 291-297 (1995).
- BERNS, E.M.J.J., KLIJN, J.G.M., VAN PUTTEN, W.L.J., VAN STAVEREN, I.L., PORTINGEN, H. and FOEKENS, J.A., *c-myc* amplification is a better prognostic factor than *HER2/neu* amplification in primary breast cancer. *Cancer Res.*, **52**, 1107-1113 (1992).
- BIÈCHE, I. and LIDEREAU, R., Genetic alterations in breast cancer. *Genes Chrom. Cancer*, **14**, 227-251 (1995).
- BORG, A., BALDETORP, B., FERNO, M., OLSSON, H. and SIGURDSSON, H., *c-myc* amplification is an independent prognostic factor in post-menopausal breast cancer. *Int. J. Cancer*, **51**, 687-691 (1992).
- CELI, F.S., COHEN, M.M., ANTONARAKIS, S.E., WERTHEIMER, E., ROTH, J. and SHULDINER, A.R., Determination of gene dosage by a quantitative adaptation of the polymerase chain reaction (qd-PCR): rapid detection of deletions and duplications of gene sequences. *Genomics*, **21**, 304-310 (1994).
- COURJAL, F., CUNY, M., SIMONY-LAFONTAINE, J., LOUASSON, G., SPEISER, P., ZEILLINGER, R., RODRIGUEZ, C. and THEILLET, C., Mapping of DNA amplifications at 15 chromosomal localizations in 1875 breast tumors: definition of phenotypic groups. *Cancer Res.*, **57**, 4360-4367 (1997).
- DENG, G., YU, M., CHEN, L.C., MOORE, D., KURISU, W., KALLIONIEMI, A., WALDMAN, F.M., COLLINS, C. and SMITH, H.S., Amplifications of oncogene *erbB-2* and chromosome 20q in breast cancer determined by differentially competitive polymerase chain reaction. *Breast Cancer Res. Treat.*, **40**, 271-281 (1996).
- GELMINI, S., ORLANDO, C., SESTINI, R., VONA, G., PINZANI, P., RUOCCO, L. and PAZZAGLI, M., Quantitative polymerase chain reaction-based homogeneous assay with fluorogenic probes to measure *c-erbB-2* oncogene amplification. *Clin. Chem.*, **43**, 752-758 (1997).
- GIBSON, U.E.M., HEID, C.A. and WILLIAMS, P.M., A novel method for real-time quantitative RT-PCR. *Genome Res.*, **6**, 995-1001 (1996).
- HEID, C.A., STEVENS, J., LIVAK, K.J. and WILLIAMS, P.M., Real-time quantitative PCR. *Genome Res.*, **6**, 986-994 (1996).
- HOLLAND, P.M., ABRAMSON, R.D., WATSON, R. and GELFAND, D.H., Detection of specific polymerase chain reaction product by utilizing the 5' to 3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. nat. Acad. Sci. (Wash.)*, **88**, 7276-7280 (1991).
- KALLIONIEMI, A., KALLIONIEMI, O.P., PIPER, J., TANNER, M., STOKKES, T., CHEN, L., SMITH, H.S., PINKEL, D., GRAY, J.W. and WALDMAN, F.M., Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc. nat. Acad. Sci. (Wash.)*, **91**, 2156-2160 (1994).
- LEE, L.G., CONNELL, C.R. and BIOCH, W., Allelic discrimination by nick-translation PCR with fluorogenic probe. *Nucleic Acids Res.*, **21**, 3761-3766 (1993).
- LONGO, N., BERNINGER, N.S. and HARTLEY, J.L., Use of uracil DNA glycosylase to control carry-over contamination in polymerase chain reactions. *Gene*, **93**, 125-128 (1990).
- MUSS, H.B., THOR, A.D., BERRY, D.A., KUTE, T., LIU, E.T., KOERNER, F., CIRINCIONE, C.T., BUDMAN, D.R., WOOD, W.C., BARCOS, M. and HENDERSON, I.C., *c-erbB-2* expression and response to adjuvant therapy in women with node-positive early breast cancer. *New Engl. J. Med.*, **330**, 1260-1266 (1994).
- PAULETTI, G., GODOLPHIN, W., PRESS, M.F. and SALMON, D.J., Detection and quantification of *HER-2/neu* gene amplification in human breast cancer archival material using fluorescence *in situ* hybridization. *Oncogene*, **13**, 63-72 (1996).
- PIATAK, M., LUK, K.C., WILLIAMS, B. and LIFSON, J.D., Quantitative competitive polymerase chain reaction for accurate quantitation of HIV DNA and RNA species. *Biotechniques*, **14**, 70-80 (1993).
- SCHUURING, E., VERHOEVEN, E., VAN TINTEREN, H., PETERSE, J.L., NUNNIK, B., THUNNISSEN, F.B.J.M., DEVILLE, P., CORNELISSE, C.J., VAN DE VIJVER, M.J., MOOI, W.J. and MICHALIDES, R.J.A.M., Amplification of genes within the chromosome 11q13 region is indicative of poor prognosis in patients with operable breast cancer. *Cancer Res.*, **52**, 5229-5234 (1992).
- SLAMON, D.J., CLARK, G.M., WONG, S.G., LEVIN, W.S., ULLRICH, A. and MCGUIRE, W.L., Human breast cancer: correlation of relapse and survival with amplification of the *HER-2/neu* oncogene. *Science*, **235**, 177-182 (1987).
- SLAMON, D.J., GODOLPHIN, W., JONES, L.A., HOLT, J.A., WONG, S.G., KEITH, D.E., LEVIN, W.J., STUART, S.G., UDOLFE, J., ULLRICH, A. and PRESS, M.F., Studies of the *HER-2/neu* proto-oncogene in human breast and ovarian cancer. *Science*, **244**, 707-712 (1989).
- VALERON, P.F., CHIRINO, R., FERNANDEZ, L., TORRES, S., NAVARRO, D., AGUIAR, J., CABRERA, J.J., DIAZ-CHICO, B.N. and DIAZ-CHICO, J.C., Validation of a differential PCR and an ELISA procedure in studying *HER-2/neu* status in breast cancer. *Int. J. Cancer*, **65**, 129-133 (1996).



# Genome-wide Study of Gene Copy Numbers, Transcripts, and Protein Levels in Pairs of Non-invasive and Invasive Human Transitional Cell Carcinomas\*

Torben F. Ørntoft<sup>‡§</sup>, Thomas Thykjaer<sup>¶</sup>, Frederic M. Waldman<sup>||</sup>, Hans Wolf<sup>\*\*</sup>, and Julio E. Celis<sup>‡‡</sup>

Gain and loss of chromosomal material is characteristic of bladder cancer, as well as malignant transformation in general. The consequences of these changes at both the transcription and translation levels is at present unknown partly because of technical limitations. Here we have attempted to address this question in pairs of non-invasive and invasive human bladder tumors using a combination of technology that included comparative genomic hybridization, high density oligonucleotide array-based monitoring of transcript levels (5600 genes), and high resolution two-dimensional gel electrophoresis. The results showed that there is a gene dosage effect that in some cases superimposes on other regulatory mechanisms. This effect depended ( $p < 0.015$ ) on the magnitude of the comparative genomic hybridization change. In general (18 of 23 cases), chromosomal areas with more than 2-fold gain of DNA showed a corresponding increase in mRNA transcripts. Areas with loss of DNA, on the other hand, showed either reduced or unaltered transcript levels. Because most proteins resolved by two-dimensional gels are unknown it was only possible to compare mRNA and protein alterations in relatively few cases of well focused abundant proteins. With few exceptions we found a good correlation ( $p < 0.005$ ) between transcript alterations and protein levels. The implications, as well as limitations, of the approach are discussed. *Molecular & Cellular Proteomics* 1:37–45, 2002.

Aneuploidy is a common feature of most human cancers (1), but little is known about the genome-wide effect of this

From the <sup>‡</sup>Department of Clinical Biochemistry, Molecular Diagnostic Laboratory and <sup>\*\*</sup>Department of Urology, Aarhus University Hospital, Skejby, DK-8200 Aarhus N, Denmark, <sup>¶</sup>AAROS Applied Biotechnology ApS, Gustav Wiedsvej 10, DK-8000 Aarhus C, Denmark, <sup>||</sup>UCSF Cancer Center and Department of Laboratory Medicine, University of California, San Francisco, CA 94143-0808, and <sup>‡‡</sup>Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Ole Worms Allé 170, Aarhus University, DK-8000 Aarhus C, Denmark

Received, September 26, 2001, and in revised form, November 7, 2001

Published, MCP Papers in Press, November 13, 2001, DOI 10.1074/mcp.M100019-MCP200

phenomenon at both the transcription and translation levels. High throughput array studies of the breast cancer cell line BT474 has suggested that there is a correlation between DNA copy numbers and gene expression in highly amplified areas (2), and studies of individual genes in solid tumors have revealed a good correlation between gene dose and mRNA or protein levels in the case of c-erb-B2, cyclin D1, *ems1*, and N-myc (3–5). However, a high cyclin D1 protein expression has been observed without simultaneous amplification (4), and a low level of c-myc copy number increase was observed without concomitant c-myc protein overexpression (6).

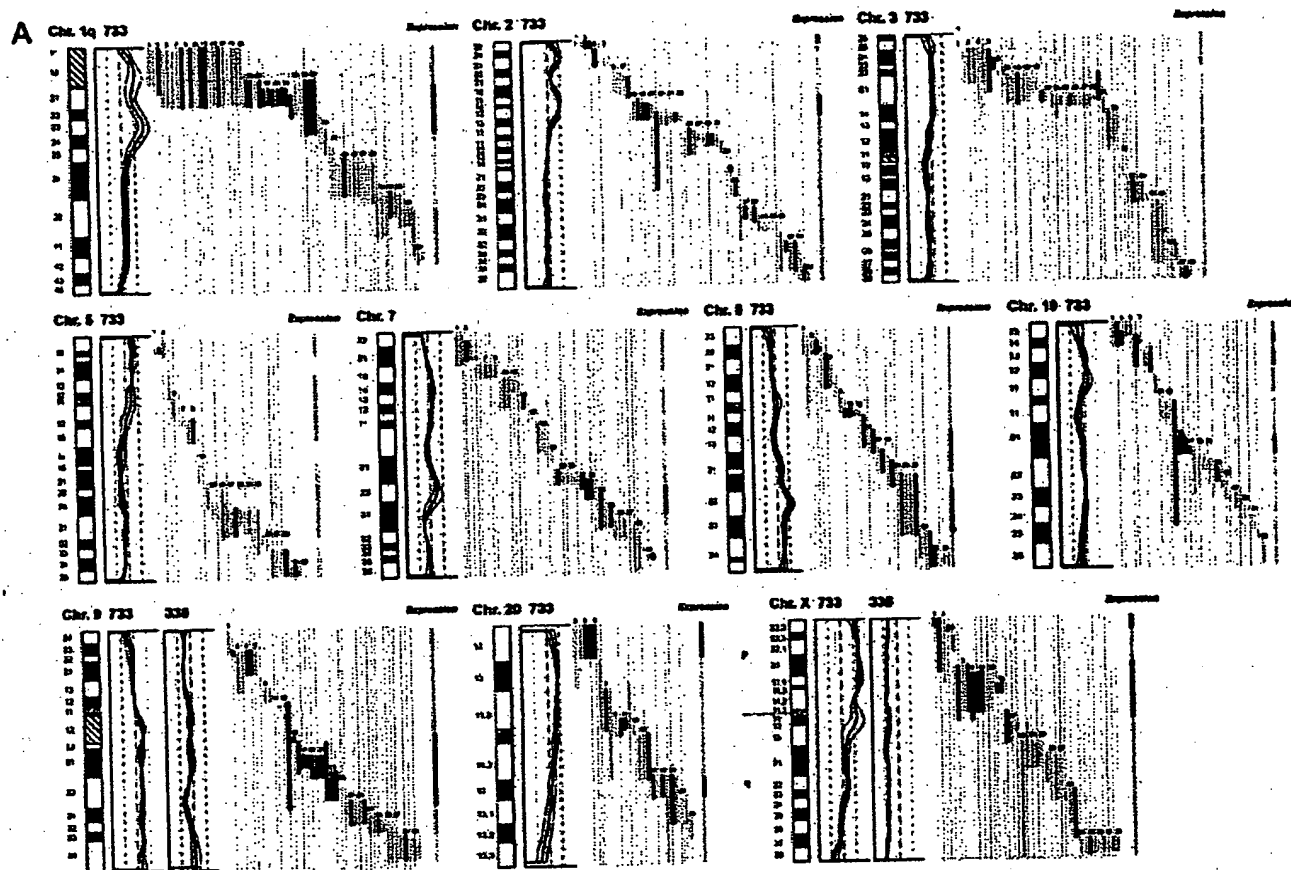
In human bladder tumors, karyotyping, fluorescent *in situ* hybridization, and comparative genomic hybridization (CGH)<sup>†</sup> have revealed chromosomal aberrations that seem to be characteristic of certain stages of disease progression. In the case of non-invasive pTa transitional cell carcinomas (TCCs), this includes loss of chromosome 9 or parts of it, as well as loss of Y in males. In minimally invasive pT1 TCCs, the following alterations have been reported: 2q–, 11p–, 1q+, 11q13+, 17q+, and 20q+ (7–12). It has been suggested that these regions harbor tumor suppressor genes and oncogenes; however, the large chromosomal areas involved often contain many genes, making meaningful predictions of the functional consequences of losses and gains very difficult.

In this investigation we have combined genome-wide technology for detecting genomic gains and losses (CGH) with gene expression profiling techniques (microarrays and proteomics) to determine the effect of gene copy number on transcript and protein levels in pairs of non-invasive and invasive human bladder TCCs.

## EXPERIMENTAL PROCEDURES

**Material**—Bladder tumor biopsies were sampled after informed consent was obtained and after removal of tissue for routine pathology examination. By light microscopy tumors 335 and 532 were staged by an experienced pathologist as pTa (superficial papillary),

<sup>†</sup> The abbreviations used are: CGH, comparative genomic hybridization; TCC, transitional cell carcinoma; LOH, loss of heterozygosity; PA-FABP, psoriasis-associated fatty acid-binding protein; 2D, two-dimensional.



**Fig. 1. DNA copy number and mRNA expression level.** Shown from left to right are chromosome (Chr.), CGH profiles, gene location and expression level of specific genes, and overall expression level along the chromosome. A, expression of mRNA in invasive tumor 733 as compared with the non-invasive counterpart tumor 335. B, expression of mRNA in invasive tumor 827 compared with the non-invasive counterpart tumor 532. The average fluorescent signal ratio between tumor DNA and normal DNA is shown along the length of the chromosome (left). The bold curve in the ratio profile represents a mean of four chromosomes and is surrounded by thin curves indicating one standard deviation. The central vertical line (broken) indicates a ratio value of 1 (no change), and the vertical lines next to it (dotted) indicate a ratio of 0.5 (left) and 2.0 (right). In chromosomes where the non-invasive tumor 335 used for comparison showed alterations in DNA content, the ratio profile of that chromosome is shown to the right of the invasive tumor profile. The colored bars represent one gene each, identified by the running numbers above the bars (the name of the gene can be seen at [www.MDLDK/sdata.html](http://www.MDLDK/sdata.html)). The bars indicate the purported location of the gene, and the colors indicate the expression level of the gene in the invasive tumor compared with the non-invasive counterpart; >2-fold increase (black), >2-fold decrease (blue), no significant change (orange). The bar to the far right, entitled Expression shows the resulting change in expression along the chromosome; the colors indicate that at least half of the genes were up-regulated (black), at least half of the genes down-regulated (blue), or more than half of the genes are unchanged (orange). If a gene was absent in one of the samples and present in another, it was regarded as more than a 2-fold change. A 2-fold level was chosen as this corresponded to one standard deviation in a double determination of ~1800 genes. Centromeres and heterochromatic regions were excluded from data analysis.

grade I and II, respectively, tumors 733 and 827 were staged as pT1 (invasive into submucosa), 733 was staged as solid, and 827 was staged as papillary, both grade III.

**mRNA Preparation**—Tissue biopsies, obtained fresh from surgery, were embedded immediately in a sodium-guanidinium thiocyanate solution and stored at  $-80^{\circ}\text{C}$ . Total RNA was isolated using the RNeasy B RNA isolation method (WAK-Chemie Medical GmbH). poly(A)<sup>+</sup> RNA was isolated by an oligo(dT) selection step (Oligotex mRNA kit; Qiagen).

**cRNA Preparation**—1  $\mu\text{g}$  of mRNA was used as starting material. The first and second strand cDNA synthesis was performed using the SuperScript<sup>®</sup> choice system (Invitrogen) according to the manufacturer's instructions but using an oligo(dT) primer containing a T7 RNA polymerase binding site. Labeled cRNA was prepared using the ME-GAscript<sup>®</sup> *in vitro* transcription kit (Ambion). Biotin-labeled CTP and

UTP (Enzo) was used, together with unlabeled NTPs in the reaction. Following the *in vitro* transcription reaction, the unincorporated nucleotides were removed using RNeasy columns (Qiagen).

**Array Hybridization and Scanning**—Array hybridization and scanning was modified from a previous method (13). 10  $\mu\text{g}$  of cRNA was fragmented at  $94^{\circ}\text{C}$  for 35 min in buffer containing 40 mM Tris acetate, pH 8.1, 100 mM KOAc, 30 mM MgOAc. Prior to hybridization, the fragmented cRNA in a 6 $\times$  SSPE-T hybridization buffer (1 M NaCl, 10 mM Tris, pH 7.6, 0.005% Triton), was heated to  $95^{\circ}\text{C}$  for 5 min, subsequently cooled to  $40^{\circ}\text{C}$ , and loaded onto the Affymetrix probe array cartridge. The probe array was then incubated for 16 h at  $40^{\circ}\text{C}$  at constant rotation (60 rpm). The probe array was exposed to 10 washes in 6 $\times$  SSPE-T at  $25^{\circ}\text{C}$  followed by 4 washes in 0.5 $\times$  SSPE-T at  $50^{\circ}\text{C}$ . The biotinylated cRNA was stained with a streptavidin-phycoerythrin conjugate, 10  $\mu\text{g}/\text{ml}$  (Molecular Probes) in 6 $\times$  SSPE-T

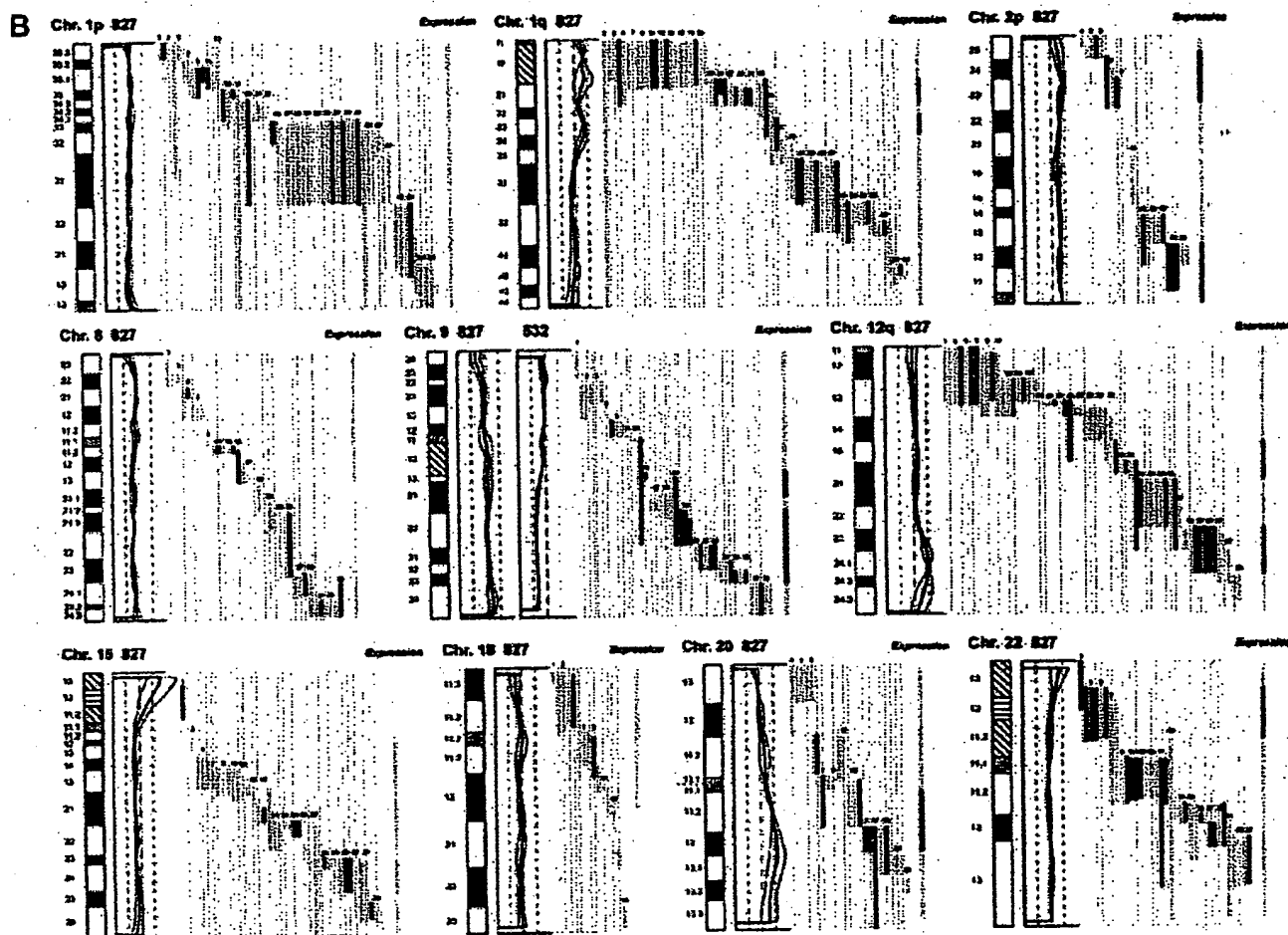


FIG. 1—continued

for 30 min at 25 °C followed by 10 washes in 6× SSPE-T at 25 °C. The probe arrays were scanned at 560 nm using a confocal laser scanning microscope (made for Affymetrix by Hewlett-Packard). The readings from the quantitative scanning were analyzed by Affymetrix gene expression analysis software.

**Microsatellite Analysis**—Microsatellite Analysis was performed as described previously (14). Microsatellites were selected by use of [www.ncbi.nlm.nih.gov/genemap98](http://www.ncbi.nlm.nih.gov/genemap98), and primer sequences were obtained from the genome data base at [www.gdb.org](http://www.gdb.org). DNA was extracted from tumor and blood and amplified by PCR in a volume of 20  $\mu$ l for 35 cycles. The amplicons were denatured and electrophoresed for 3 h in an ABI Prism 377. Data were collected in the Gene Scan program for fragment analysis. Loss of heterozygosity was defined as less than 33% of one allele detected in tumor amplicons compared with blood.

**Proteomic Analysis**—TCCs were minced into small pieces and homogenized in a small glass homogenizer in 0.5 ml of lysis solution. Samples were stored at -20 °C until use. The procedure for 2D gel electrophoresis has been described in detail elsewhere (15, 16). Gels were stained with silver nitrate and/or Coomassie Brilliant Blue. Proteins were identified by a combination of procedures that included microsequencing, mass spectrometry, two-dimensional gel Western immunoblotting, and comparison with the master two-dimensional gel image of human keratinocyte proteins; see [biobase.dk/cgi-bin/celis](http://biobase.dk/cgi-bin/celis).

**CGH**—Hybridization of differentially labeled tumor and normal DNA to normal metaphase chromosomes was performed as described previously (10). Fluorescein-labeled tumor DNA (200 ng), Texas Red-

labeled reference DNA (200 ng), and human Cot-1 DNA (20  $\mu$ g) were denatured at 37 °C for 5 min and applied to denatured normal metaphase slides. Hybridization was at 37 °C for 2 days. After washing, the slides were counterstained with 0.15  $\mu$ g/ml 4,6-diamidino-2-phenylindole in an anti-fade solution. A second hybridization was performed for all tumor samples using fluorescein-labeled reference DNA and Texas Red-labeled tumor DNA (inverse labeling) to confirm the aberrations detected during the initial hybridization. Each CGH experiment also included a normal control hybridization using fluorescein- and Texas Red-labeled normal DNA. Digital image analysis was used to identify chromosomal regions with abnormal fluorescence ratios, indicating regions of DNA gains and losses. The average green:red fluorescence intensity ratio profiles were calculated using four images of each chromosome (eight chromosomes total) with normalization of the green:red fluorescence intensity ratio for the entire metaphase and background correction. Chromosome identification was performed based on 4,6-diamidino-2-phenylindole banding patterns. Only images showing uniform high intensity fluorescence with minimal background staining were analyzed. All centromeres, p arms of acrocentric chromosomes, and heterochromatic regions were excluded from the analysis.

## RESULTS

**Comparative Genomic Hybridization**—The CGH analysis identified a number of chromosomal gains and losses in the

# Gene Copy Numbers, Transcripts, and Protein Levels

TABLE I  
Correlation between alterations detected by CGH and by expression monitoring

Top, CGH used as independent variable (if CGH alteration – what expression ratio was found); bottom, altered expression used as independent variable (if expression alteration – what CGH deviation was found).

CGH alterations	Tumor 733 vs. 335	Concordance	CGH alterations	Tumor 827 vs. 532	Concordance
	Expression change clusters			Expression change clusters	
13 Gain	10 Up-regulation 0 Down-regulation 3 No change	77%	10 Gain	8 Up-regulation 0 Down-regulation 2 No change	80%
10 Loss	1 Up-regulation 5 Down-regulation 4 No change	50%	12 Loss	3 Up-regulation 2 Down regulation 7 No change	17%
Expression change clusters	Tumor 733 vs. 335	Concordance	Expression change clusters	Tumor 827 vs. 532	Concordance
	CGH alterations			CGH alterations	
16 Up-regulation	11 Gain 2 Loss 3 No change	69%	17 Up-regulation	10 Gain 5 Loss 2 No change	59%
21 Down-regulation	1 Gain 8 Loss 12 No change	38%	9 Down-regulation	0 Gain 3 Loss 6 No change	33%
15 No change	3 Gain 3 Loss 9 No change	60%	21 No change	1 Gain 3 Loss 17 No change	81%

two invasive tumors (stage pT1, TCCs 733 and 827), whereas the two non-invasive papillomas (stage pTa, TCCs 335 and 532) showed only 9p–, 9q22–q33–, and X–, and 7+, 9q–, and Y–, respectively. Both invasive tumors showed changes (1q22–24+, 2q14.1–qter–, 3q12–q13.3–, 6q12–q22–, 9q34+, 11q12–q13+, 17+, and 20q11.2–q12+) that are typical for their disease stage, as well as additional alterations, some of which are shown in Fig. 1. Areas with gains and losses deviated from the normal copy number to some extent, and the average numerical deviation from normal was 0.4-fold in the case of TCC 733 and 0.3-fold for TCC 827. The largest changes, amounting to at least a doubling of chromosomal content, were observed at 1q23 in TCC 733 (Fig. 1A) and 20q12 in TCC 827 (Fig. 1B).

**mRNA Expression in Relation to DNA Copy Number**—The mRNA levels from the two invasive tumors (TCCs 827 and 733) were compared with the two non-invasive counterparts (TCCs 532 and 335). This was done in two separate experiments in which we compared TCCs 733 to 335 and 827 to 532, respectively, using two different scaling settings for the arrays to rule out scaling as a confounding parameter. Approximately 1,800 genes that yielded a signal on the arrays were searched in the Unigene and Genemap data bases for chromosomal location, and those with a known location (1096) were plotted as bars covering their purported locus. In that way it was possible to construct a graphic presentation of DNA copy number and relative mRNA levels along the individual chromosomes (Fig. 1).

For each mRNA a ratio was calculated between the level in the invasive versus the non-invasive counterpart. Bars, which represent chromosomal location of a gene, were color-coded according to the expression ratio, and only differences larger

than 2-fold were regarded as informative (Fig. 1). The density of genes along the chromosomes varied, and areas containing only one gene were excluded from the calculations. The resolution of the CGH method is very low, and some of the outlier data may be because of the fact that the boundaries of the chromosomal aberrations are not known at high resolution.

Two sets of calculations were made from the data. For the first set we used CGH alterations as the independent variable and estimated the frequency of expression alterations in these chromosomal areas. In general, areas with a strong gain of chromosomal material contained a cluster of genes having increased mRNA expression. For example, both chromosomes 1q21–q25, 2p and 9q, showed a relative gain of more than 100% in DNA copy number that was accompanied by increased mRNA expression levels in the two tumor pairs (Fig. 1). In most cases, chromosomal gains detected by CGH were accompanied by an increased level of transcripts in both TCCs 733 (77%) and 827 (80%) (Table I, top). Chromosomal losses, on the other hand, were not accompanied by decreased expression in several cases, and were often registered as having unaltered RNA levels (Table I, top). The inability to detect RNA expression changes in these cases was not because of fewer genes mapping to the lost regions (data not shown).

In the second set of calculations we selected expression alterations above 2-fold as the independent variable and estimated the frequency of CGH alterations in these areas. As above, we found that increased transcript expression correlated with gain of chromosomal material (TCC 733, 69% and TCC 827, 59%), whereas reduced expression was often detected in areas with unaltered CGH ratios (Table I, bottom). Furthermore, as a control we looked at areas with no alter-

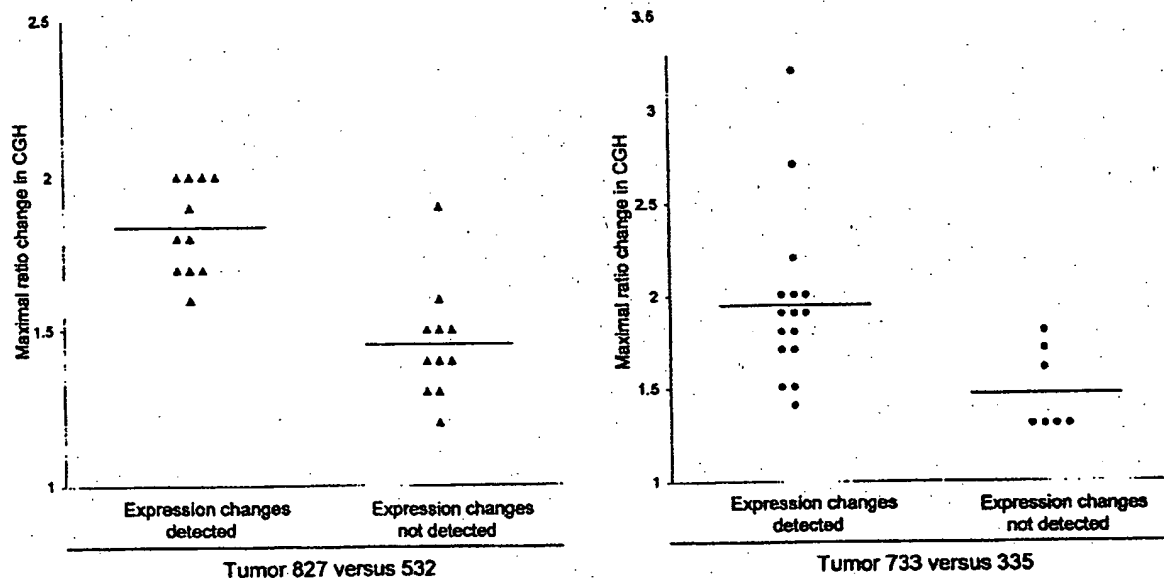


FIG. 2. Correlation between maximum CGH aberration and the ability to detect expression change by oligonucleotide array monitoring. The aberration is shown as a numerical -fold change in ratio between invasive tumors 827 (▲) and 733 (◆) and their non-invasive counterparts 532 and 335. The expression change was taken from the *Expression* line to the right in Fig. 1, which depicts the resulting expression change for a given chromosomal region. At least half of the mRNAs from a given region have to be either up- or down-regulated to be scored as an expression change. All chromosomal arms in which the CGH ratio plus or minus one standard deviation was outside the ratio value of one were included.

ation in expression. No alteration was detected by CGH in most of these areas (TCC 733, 60% and TCC 827, 81%; see Table I, bottom). Because the ability to observe reduced or increased mRNA expression clustering to a certain chromosomal area clearly reflected the extent of copy number changes, we plotted the maximum CGH aberrations in the regions showing CGH changes against the ability to detect a change in mRNA expression as monitored by the oligonucleotide arrays (Fig. 2). For both tumors TCC 733 ( $p < 0.015$ ) and TCC 827 ( $p < 0.00003$ ) a highly significant correlation was observed between the level of CGH ratio change (reflecting the DNA copy number) and alterations detected by the array based technology (Fig. 2). Similar data were obtained when areas with altered expression were used as independent variables. These areas correlated best with CGH when the CGH ratio deviated 1.6- to 2.0-fold (Table I, bottom) but mostly did not at lower CGH deviations. These data probably reflect that loss of an allele may only lead to a 50% reduction in expression level, which is at the cut-off point for detection of expression alterations. Gain of chromosomal material can occur to a much larger extent.

**Microsatellite-based Detection of Minor Areas of Losses**—In TCC 733, several chromosomal areas exhibiting DNA amplification were preceded or followed by areas with a normal CGH but reduced mRNA expression (see Fig. 1, TCC 733 chromosome 1q32, 2p21, and 7q21 and q32, 9q34, and 10q22). To determine whether these results were because of undetected loss of chromosomal material in these regions or

because of other non-structural mechanisms regulating transcription, we examined two microsatellites positioned at chromosome 1q25–32 and two at chromosome 2p22. Loss of heterozygosity (LOH) was found at both 1q25 and at 2p22 indicating that minor deleted areas were not detected with the resolution of CGH (Fig. 3). Additionally, chromosome 2p in TCC 733 showed a CGH pattern of gain/no change/gain of DNA that correlated with transcript increase/decrease/increase. Thus, for the areas showing increased expression there was a correlation with the DNA copy number alterations (Fig. 1A). As indicated above, the mRNA decrease observed in the middle of the chromosomal gain was because of LOH, implying that one of the mechanisms for mRNA down-regulation may be regions that have undergone smaller losses of chromosomal material. However, this cannot be detected with the resolution of the CGH method.

In both TCC 733 and TCC 827, the telomeric end of chromosome 11p showed a normal ratio in the CGH analysis; however, clusters of five and three genes, respectively, lost their expression. Two microsatellites (D11S1760, D11S922) positioned close to MUC2, IGF2, and cathepsin D indicated LOH as the most likely mechanism behind the loss of expression (data not shown).

A reduced expression of mRNA observed in TCC 733 at chromosomes 3q24, 11p11, 12p12.2, 12q21.1, and 16q24 and in TCC 827 at chromosome 11p15.5, 12p11, 15q11.2, and 18q12 was also examined for chromosomal losses using microsatellites positioned as close as possible to the gene loci

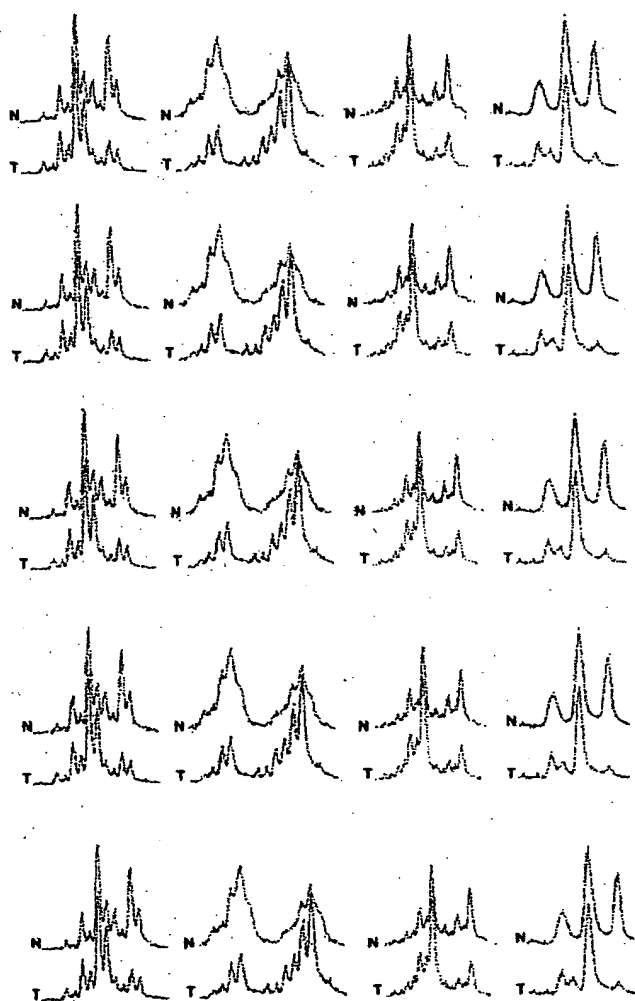


FIG. 3. Microsatellite analysis of loss of heterozygosity. Tumor 733 showing loss of heterozygosity at chromosome 1q25, detected (a) by D1S215 close to Hu class I histocompatibility antigen (gene number 38 in Fig. 1), (b) by D1S2735 close to cathepsin E (gene number 41 in Fig. 1), and (c) at chromosome 2p23 by D2S2251 close to general  $\beta$ -spectrin (gene number 11 on Fig. 1) and of (d) tumor 827 showing loss of heterozygosity at chromosome 18q12 by S18S1118 close to mitochondrial 3-oxoacyl-coenzyme A thiolase (gene number 12 in Fig. 1). The upper curves show the electropherogram obtained from normal DNA from leukocytes (N), and the lower curves show the electropherogram from tumor DNA (T). In all cases one allele is partially lost in the tumor amplicon.

showing reduced mRNA transcripts. Only the microsatellite positioned at 18q12 showed LOH (Fig. 3), suggesting that transcriptional down-regulation of genes in the other regions may be controlled by other mechanisms.

**Relation between Changes in mRNA and Protein Levels—**2D-PAGE analysis, in combination with Coomassie Brilliant Blue and/or silver staining, was carried out on all four tumors using fresh biopsy material. 40 well resolved abundant known proteins migrating in areas away from the edges of the pH

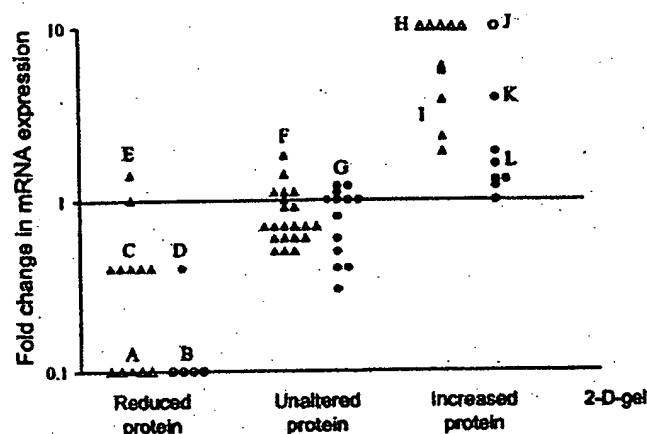


FIG. 4. Correlation between protein levels as judged by 2D-PAGE and transcript ratio. For comparison proteins were divided in three groups, unaltered in level or up- or down-regulated (horizontal axis). The mRNA ratio as determined by oligonucleotide arrays was plotted for each gene (vertical axis).  $\Delta$ , mRNAs that were scored as present in both tumors used for the ratio calculation;  $\Delta$ , mRNAs that were scored as absent in the invasive tumors (along horizontal axis) or as absent in non-invasive reference (top of figure). Two different scalings were used to exclude scaling as a confounder, TCCs 827 and 532 ( $\Delta\Delta$ ) were scaled with background suppression, and TCCs 733 and 335 ( $\bullet\bullet$ ) were scaled without suppression. Both comparisons showed highly significant ( $p < 0.005$ ) differences in mRNA ratios between the groups. Proteins shown were as follows: Group A (from left), phosphoglucosylase 1, glutathione transferase class  $\mu$  number 4, fatty acid-binding protein homologue, cytochrome 15, and cytochrome 13; B (from left), fatty acid-binding protein homologue, 28-kDa heat shock protein, cytochrome 13, and calnexin; C (from left),  $\alpha$ -enolase, hnRNP B1, 28-kDa heat shock protein, 14-3-3- $\epsilon$ , and pre-mRNA splicing factor; D, mesothelial keratin K7 (type II); E (from top), glutathione S-transferase- $\pi$  and mesothelial keratin K7 (type II); F (from top and left), adenyl cyclase-associated protein, E-cadherin, keratin 19, calgizzarin, phosphoglycerate mutase, annexin IV, cytoskeletal  $\gamma$ -actin, hnRNP A1, integral membrane protein calnexin (IP90), hnRNP H, brain-type clathrin light chain- $\alpha$ , hnRNP F, 70-kDa heat shock protein, heterogeneous nuclear ribonucleoprotein A/B, translationally controlled tumor protein, liver glyceraldehyde-3-phosphate dehydrogenase, keratin 8, aldehyde reductase, and Na,K-ATPase  $\beta$ -1 subunit; G, (from top and left), TCP20, calgizzarin, 70-kDa heat shock protein, calnexin, hnRNP H, cytochrome 15, ATP synthase, keratin 19, triosephosphate isomerase, hnRNP F, liver glyceraldehyde-3-phosphate dehydrogenase, glutathione S-transferase- $\pi$ , and keratin 8; H (from left), plasma gelsolin, autoantigen calreticulin, thioredoxin, and NAD $^{+}$ -dependent 15 hydroxyprostaglandin dehydrogenase; I (from top), prollyl 4-hydroxylase  $\beta$ -subunit, cytochrome 20, cytochrome 17, prothionin, and fructose 1,6-bisphosphatase; J annexin II; K, annexin IV; L (from top and left), 90-kDa heat shock protein, prollyl 4-hydroxylase  $\beta$ -subunit,  $\alpha$ -enolase, GRP 78, cyclophilin, and cofilin.

gradient, and having a known chromosomal location, were selected for analysis in the TCC pair 827/532. Proteins were identified by a combination of methods (see "Experimental Procedures"). In general there was a highly significant correlation ( $p < 0.005$ ) between mRNA and protein alterations (Fig. 4). Only one gene showed disagreement between transcript alteration and protein alteration. Except for a group of cyto-

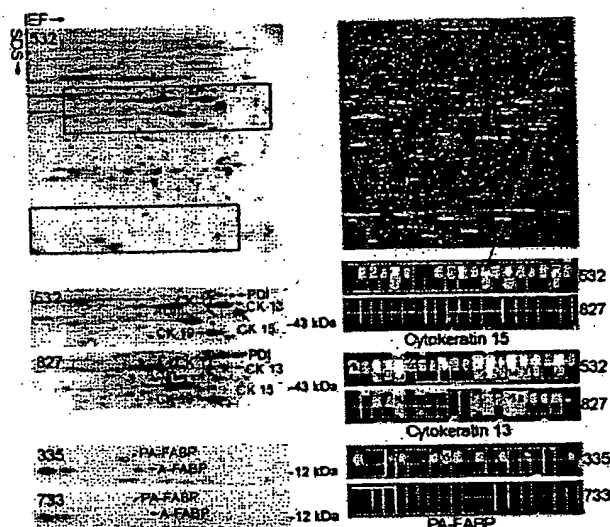


FIG. 5. Comparison of protein and transcript levels in invasive and non-invasive TCCs. The upper part of the figure shows a 2D gel (left) and the oligonucleotide array (right) of TCC 532. The red rectangles on the upper gel highlight the areas that are compared below. Identical areas of 2D gels of TCCs 532 and 827 are shown below. Clearly, cytokeratins 13 and 15 are strongly down-regulated in TCC 827 (red annotation). The tile on the array containing probes for cytokeratin 15 is enlarged below the array (red arrow) from TCC 532 and is compared to perfect match probes; the lower row corresponds to mismatch probes containing a mutation (used for correction for unspecific binding). Absence of signal is depicted as black, and the higher the signal the lighter the color. A high transcript level was detected in TCC 532 (6151 units) whereas a much lower level was detected in TCC 827 (absence of signals). For cytokeratin 13, a high transcript level was also present in TCC 532 (15659 units), and a much lower level was present in TCC 827 (623 units). The 2D gels at the bottom of the figure (left) show levels of PA-FABP and adipocyte-FABP in TCCs 335 and 733 (invasive), respectively. Both proteins are down-regulated in the invasive tumor. To the right we show the array tiles for the PA-FABP transcript. A medium transcript level was detected in the case of TCC 335 (1277 units) whereas very low levels were detected in TCC 733 (166 units). IEF, isoelectric focusing.

keratins encoded by genes on chromosome 17 (Fig. 5) the analyzed proteins did not belong to a particular family. 26 well focused proteins whose genes had a known chromosomal location were detected in TCCs 733 and 335, and of these 19 correlated ( $p < 0.005$ ) with the mRNA changes detected using the arrays (Fig. 4). For example, PA-FABP was highly expressed in the non-invasive TCC 335 but lost in the invasive counterpart (TCC 733; see Fig. 5). The smaller number of proteins detected in both 733 and 335 was because of the smaller size of the biopsies that were available.

11 chromosomal regions where CGH showed aberrations that corresponded to the changes in transcript levels also showed corresponding changes in the protein level (Table II). These regions included genes that encode proteins that are found to be frequently altered in bladder cancer, namely cytokeratins 17 and 20, annexins II and IV, and the fatty acid-binding proteins PA-FABP and FABP1. Four of these proteins were encoded by genes in chromosome 17q, a frequently amplified chromosomal area in invasive bladder cancers.

#### DISCUSSION

Most human cancers have abnormal DNA content, having lost some chromosomal parts and gained others. The present study provides some evidence as to the effect of these gains and losses on gene expression in two pairs of non-invasive and invasive TCCs using high throughput expression arrays and proteomics, in combination with CGH. In general, the results showed that there is a clear individual regulation of the mRNA expression of single genes, which in some cases was superimposed by a DNA copy number effect. In most cases, genes located in chromosomal areas with gains often exhibited increased mRNA expression, whereas areas showing losses showed either no change or a reduced mRNA expression. The latter might be because of the fact that losses most often are restricted to loss of one allele, and the cut-off point for detection of expression alterations was a 2-fold change, thus being at the border of detection. In several cases, how-

TABLE II  
Proteins whose expression level correlates with both mRNA and gene dose changes

Protein	Chromosomal location	Tumor TCC	CGH alteration	Transcript alteration <sup>a</sup>	Protein alteration
Annexin II	1q21	733	Gain	Abs to Pres <sup>a</sup>	Increase
Annexin IV	2p13	733	Gain	3.9-Fold up	Increase
Cytokeratin 17	17q12-q21	827	Gain	3.8-Fold up	Increase
Cytokeratin 20	17q21.1	827	Gain	5.6-Fold up	Increase
(PA-)FABP	8q21.2	827	Loss	10-Fold down	Decrease
FABP1	9q22	827	Gain	2.3-Fold up	Increase
Plasma gelsolin	9q31	827	Gain	Abs to Pres	Increase
Heat shock protein 28	15q12-q13	827	Loss	2.5-Fold up	Decrease
Prohibitin	17q21	827/733	Gain	3.7-/2.5-Fold up <sup>b</sup>	Increase
Prolyl-4-hydroxyl	17q25	827/733	Gain	5.7-/1.6-Fold up	Increase
hnRNPI	7p15	827	Loss	2.5-Fold down	Decrease

<sup>a</sup> Abs, absent; Pres, present.

<sup>b</sup> In cases where the corresponding alterations were found in both TCCs 827 and 733 these are shown as 827/733.



ever, an increase or decrease in DNA copy number was associated with *de novo* occurrence or complete loss of transcript, respectively. Some of these transcripts could not be detected in the non-invasive tumor but were present at relatively high levels in areas with DNA amplifications in the invasive tumors (e.g. in TCC 733 transcript from cellular ligand of annexin II gene (chromosome 1q21) from absent to 2670 arbitrary units; in TCC 827 transcript from small proline-rich protein 1 gene (chromosome 1q12-q21.1) from absent to 1326 arbitrary units). It may be anticipated from these data that significant clustering of genes with an increased expression to a certain chromosomal area indicates an increased likelihood of gain of chromosomal material in this area.

Considering the many possible regulatory mechanisms acting at the level of transcription, it seems striking that the gene dose effects were so clearly detectable in gained areas. One hypothetical explanation may lie in the loss of controlled methylation in tumor cells (17-19). Thus, it may be possible that in chromosomes with increased DNA copy numbers two or more alleles could be demethylated simultaneously leading to a higher transcription level, whereas in chromosomes with losses the remaining allele could be partly methylated, turning off the process (20, 21). A recent report has documented a ploidy regulation of gene expression in yeast, but in this case all the genes were present in the same ratio (22), a situation that is not analogous to that of cancer cells, which show marked chromosomal aberrations, as well as gene dosage effects.

Several CGH studies of bladder cancer have shown that some chromosomal aberrations are common at certain stages of disease progression, often occurring in more than 1 of 3 tumors. In pTa tumors, these include 9p-, 9q-, 1q+, Y- (2, 6), and in pT1 tumors, 2q-, 11p-, 11q-, 1q+, 5p+, 8q+, 17q+, and 20q+ (2-4, 6, 7). The pTa tumors studied here showed similar aberrations such as 9p- and 9q22-q33- and 9q- and Y-, respectively. Likewise, the two minimal invasive pT1 tumors showed aberrations that are commonly seen at that stage, and TCC 827 had a remarkable resemblance to the commonly seen pattern of losses and gains, such as 1q22-24 amplification (seen in both tumors), 11q14-q22 loss, the latter often linked to 17 q+ (both tumors), and 1q+ and 9p-, often linked to 20q+ and 11 q13+ (both tumors) (7-9). These observations indicate that the pairs of tumors used in this study exhibit chromosomal changes observed in many tumors, and therefore the findings could be of general importance for bladder cancer.

Considering that the mapping resolution of CGH is of about 20 megabases it is only possible to get a crude picture of chromosomal instability using this technique. Occasionally, we observed reduced transcript levels close to or inside regions with increased copy numbers. Analysis of these regions by positioning heterozygous microsatellites as close as possible to the locus showing reduced gene expression revealed loss of heterozygosity in several cases. It seems likely that multiple and different events occur along each chromosomal

arm and that the use of cDNA microarrays for analysis of DNA copy number changes will reach a resolution that can resolve these changes, as has recently been proposed (2). The outlier data were not more frequent at the boundaries of the CGH aberrations. At present we do not know the mechanism behind chromosomal aneuploidy and cannot predict whether chromosomal gains will be transcribed to a larger extent than the two native alleles. A mechanism as genetic imprinting has an impact on the expression level in normal cells and is often reduced in tumors. However, the relation between imprinting and gain of chromosomal material is not known.

We regard it as a strength of this investigation that we were able to compare invasive tumors to benign tumors rather than to normal urothelium, as the tumors studied were biologically very close and probably may represent successive steps in the progression of bladder cancer. Despite the limited amount of fresh tissue available it was possible to apply three different state of the art methods. The observed correlation between DNA copy number and mRNA expression is remarkable when one considers that different pieces of the tumor biopsies were used for the different sets of experiments. This indicates that bladder tumors are relatively homogenous, a notion recently supported by CGH and LOH data that showed a remarkable similarity even between tumors and distant metastasis (10, 23).

In the few cases analyzed, mRNA and protein levels showed a striking correspondence although in some cases we found discrepancies that may be attributed to translational regulation, post-translational processing, protein degradation, or a combination of these. Some transcripts belong to undertranslated mRNA pools, which are associated with few translationally inactive ribosomes; these pools, however, seem to be rare (24). Protein degradation, for example, may be very important in the case of polypeptides with a short half-life (e.g. signaling proteins). A poor correlation between mRNA and protein levels was found in liver cells as determined by arrays and 2D-PAGE (25), and a moderate correlation was recently reported by Ideker *et al.* (26) in yeast.

Interestingly, our study revealed a much better correlation between gained chromosomal areas and increased mRNA levels than between loss of chromosomal areas and reduced mRNA levels. In general, the level of CGH change determined the ability to detect a change in transcript. One possible explanation could be that by losing one allele the change in mRNA level is not so dramatic as compared with gain of material, which can be rather unlimited and may lead to a severalfold increase in gene copy number resulting in a much higher impact on transcript level. The latter would be much easier to detect on the expression arrays as the cut-off point was placed at a 2-fold level so as not to be biased by noise on the array. Construction of arrays with a better signal to noise ratio may in the future allow detection of lesser than 2-fold alterations in transcript levels, a feature that may facilitate the analysis of the effect of loss of chromosomal areas on transcript levels.



In eleven cases we found a significant correlation between DNA copy number, mRNA expression, and protein level. Four of these proteins were encoded by genes located at a frequently amplified area in chromosome 17q. Whether DNA copy number is one of the mechanisms behind alteration of these eleven proteins is at present unknown and will have to be proved by other methods using a larger number of samples. One factor making such studies complicated is the large extent of protein modification that occurs after translation, requiring immunoidentification and/or mass spectrometry to correctly identify the proteins in the gels.

In conclusion, the results presented in this study exemplify the large body of knowledge that may be possible to gather in the future by combining state of the art techniques that follow the pathway from DNA to protein (26). Here, we used a traditional chromosomal CGH method, but in the future high resolution CGH based on microarrays with many thousand radiation hybrid-mapped genes will increase the resolution and information derived from these types of experiments (2). Combined with expression arrays analyzing transcripts derived from genes with known locations, and 2D gel analysis to obtain information at the post-translational level, a clearer and more developed understanding of the tumor genome will be forthcoming.

**Acknowledgments**—We thank Mie Madsen, Hanne Steen, Inge Lis Thorsen, Hans Lund, Vikolaj Ørntoft, and Lynn Bjerke for technical help and Thomas Gingeras, Christine Harrington, and Morten Østergaard for valuable discussions.

\* This work was supported by grants from The Danish Cancer Society, the University of Aarhus, Aarhus County, Novo Nordic, the Danish Biotechnology Program, the Frenkels Foundation, the John and Birthe Meyer Foundation, and NCI, National Institutes of Health Grant CA47537. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed: Dept. of Clinical Biochemistry, Molecular Diagnostic Laboratory, Aarhus University Hospital, Skejby, DK-8200 Aarhus N, Denmark. Tel.: 45-89495100/45-86156201 (private); Fax: 45-89496018; E-mail: orntoft@kba.sks.au.dk.

## REFERENCES

- Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998) Genetic instabilities in human cancers. *Nature* 396, 643–649.
- Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23, 41–46.
- de Cremoux, P., Martin, E. C., Vincent-Salomon, A., Dieras, V., Barbaroux, C., Liva, S., Pouillart, P., Sastre-Garau, X., and Magdelenat, H. (1999) Quantitative PCR analysis of c-erb B-2 (HER2/neu) gene amplification and comparison with p185(HER2/neu) protein expression in breast cancer drill biopsies. *Int. J. Cancer* 83, 157–161.
- Brugier, P. P., Tamimi, Y., Shuuring, E., and Schalken, J. (1996) Expression of cyclin D1 and EMS1 in bladder tumors; relationship with chromosome 11q13 amplifications. *Oncogene* 12, 1747–1753.
- Slavc, I., Ellenbogen, R., Jung, W. H., Vawter, G. F., Kretschmar, C., Grier, H., and Korf, B. R. (1990) *myc* gene amplification and expression in primary human neuroblastoma. *Cancer Res.* 50, 1459–1463.
- Sauter, G., Carroll, P., Moch, H., Kallioniemi, A., Kerschmann, R., Narayan, P., Mihatsch, M. J., and Waldman, F. M. (1995) *c-myc* copy number gains in bladder cancer detected by fluorescence *in situ* hybridization. *Am. J. Pathol.* 146, 1131–1139.
- Richter, J., Jiang, F., Gorog, J. P., Sartorius, G., Egenter, C., Gasser, T. C., Moch, H., Mihatsch, M. J., and Sauter, G. (1997) Marked genetic differences between stage pTa and stage pT1 papillary bladder cancer detected by comparative genomic hybridization. *Cancer Res.* 57, 2860–2864.
- Richter, J., Beffa, L., Wagner, U., Schraml, P., Gasser, T. C., Moch, H., Mihatsch, M. J., and Sauter, G. (1998) Patterns of chromosomal imbalances in advanced urinary bladder cancer detected by comparative genomic hybridization. *Am. J. Pathol.* 153, 1615–1621.
- Bruch, J., Wöhr, G., Hautmann, R., Mattfeldt, T., Bruderlein, S., Möller, P., Sauter, S., Hameister, H., Vogel, W., and Paiss, T. (1998) Chromosomal changes during progression of transitional cell carcinoma of the bladder and delineation of the amplified interval on chromosome arm 8q. *Genes Chromosomes Cancer* 23, 167–174.
- Hovey, R. M., Chu, L., Balazs, M., De Vries, S., Moore, D., Sauter, G., Carroll, P. R., and Waldman, F. M. (1998) Genetic alterations in primary bladder cancers and their metastases. *Cancer Res.* 58, 3555–3560.
- Simon, R., Burger, H., Brinkschmidt, C., Bocker, W., Hertle, L., and Terpe, H. J. (1998) Chromosomal aberrations associated with invasion in papillary superficial bladder cancer. *J. Pathol.* 185, 345–351.
- Koo, S. H., Kwon, K. C., Ihm, C. H., Jeon, Y. M., Park, J. W., and Sul, C. K. (1999) Detection of genetic alterations in bladder tumors by comparative genomic hybridization and cytogenetic analysis. *Cancer Genet. Cytogenet.* 110, 87–93.
- Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15, 1359–1367.
- Christensen, M., Sunde, L., Bolund, L., and Ørntoft, T. F. (1999) Comparison of three methods of microsatellite detection. *Scand. J. Clin. Lab. Invest.* 59, 167–177.
- Celis, J. E., Østergaard, M., Basse, B., Celis, A., Lauridsen, J. B., Ratz, G. P., Andersen, I., Hein, B., Wolf, H., Ørntoft, T. F., and Rasmussen, H. H. (1996) Loss of adipocyte-type fatty acid binding protein and other protein biomarkers is associated with progression of human bladder transitional cell carcinomas. *Cancer Res.* 56, 4782–4790.
- Celis, J. E., Ratz, G., Basse, B., Lauridsen, J. B., and Celis, A. (1994) In *Cell Biology: A Laboratory Handbook* (Cells, J. E., ed) Vol. 3, pp. 222–230, Academic Press, Orlando, FL.
- Ohlsson, R., Tycko, B., and Sapientza, C. (1998) Monoallelic expression: 'there can only be one'. *Trends Genet.* 14, 435–438.
- Hollander, G. A., Zuklys, S., Morel, C., Mizoguchi, E., Mobisson, K., Simpson, S., Terhorst, C., Wishart, W., Golani, D. E., Bhan, A. K., and Burakoff, S. J. (1998) Monoallelic expression of the interleukin-2 locus. *Science* 279, 2118–2121.
- Brannan, C. I., and Bartolomei, M. S. (1999) Mechanisms of genomic imprinting. *Curr. Opin. Genet. Dev.* 9, 164–170.
- Ohlsson, R., Cui, H., He, L., Pfeifer, S., Malmikumpu, H., Jiang, S., Feinberg, A. P., and Hedborg, F. (1999) Mosaic allelic insulin-like growth factor 2 expression patterns reveal a link between Wilms' tumorigenesis and epigenetic heterogeneity. *Cancer Res.* 59, 3889–3892.
- Cui, H., Hedborg, F., He, L., Nordenskjöld, A., Sandstedt, B., Pfeifer-Ohlsson, S., and Ohlsson, R. (1997) Inactivation of H19, an imprinted and putative tumor repressor gene, is a preneoplastic event during Wilms' tumorigenesis. *Cancer Res.* 57, 4469–4473.
- Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S., and Fink, G. R. (1999) Ploidy regulation of gene expression. *Science* 285, 251–254.
- Tsao, J., Yatabe, Y., Markl, I. D., Haiyan, K., Jones, P. A., and Shibata, D. (2000) Bladder cancer genotype stability during clinical progression. *Genes Chromosomes Cancer* 28, 26–32.
- Zong, Q., Schummer, M., Hood, L., and Morris, D. R. (1999) Messenger RNA translation state: the second dimension of high-throughput expression screening. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10632–10636.
- Anderson, L., and Seilhamer, J. (1997) Comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537.
- Ideker, T., Thorsson, V., Raniish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001) Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science* 292, 929–934.

# Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer<sup>1,2</sup>

Elizabeth Hyman,<sup>3</sup> Päivikki Kauraniemi,<sup>3</sup> Sampsa Hautaniemi, Maija Wolf, Spyro Mousses, Ester Rozenblum, Markus Ringnér, Guido Sauter, Outi Monni, Abdel Elkahoul, Olli-P. Kallioniemi, and Anne Kallioniemi<sup>4</sup>

Howard Hughes Medical Institute-NIH Research Scholar, Bethesda, Maryland 20892 [E.H.]; Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892 [E.H., P.K., S.H., M.W., S.M., E.R., M.R., A.E., O.K., A.K.]; Laboratory of Cancer Genetics, Institute of Medical Technology, University of Tampere and Tampere University Hospital, FIN-33520 Tampere, Finland [P.K., A.K.]; Signal Processing Laboratory, Tampere University of Technology, FIN-33101 Tampere, Finland [S.H.]; Institute of Pathology, University of Basel, CH-4003 Basel, Switzerland [G.S.]; and Biomedicum Biochip Center, Helsinki University Hospital, Biomedicum Helsinki, FIN-00014 Helsinki, Finland [O.M.]

## ABSTRACT

Genetic changes underlie tumor progression and may lead to cancer-specific expression of critical genes. Over 1100 publications have described the use of comparative genomic hybridization (CGH) to analyze the pattern of copy number alterations in cancer, but very few of the genes affected are known. Here, we performed high-resolution CGH analysis on cDNA microarrays in breast cancer and directly compared copy number and mRNA expression levels of 13,824 genes to quantitate the impact of genomic changes on gene expression. We identified and mapped the boundaries of 24 independent amplicons, ranging in size from 0.2 to 12 Mb. Throughout the genome, both high- and low-level copy number changes had a substantial impact on gene expression, with 44% of the highly amplified genes showing overexpression and 10.5% of the highly overexpressed genes being amplified. Statistical analysis with random permutation tests identified 270 genes whose expression levels across 14 samples were systematically attributable to gene amplification. These included most previously described amplified genes in breast cancer and many novel targets for genomic alterations, including the *HOXB7* gene, the presence of which in a novel amplicon at 17q21.3 was validated in 10.2% of primary breast cancers and associated with poor patient prognosis. In conclusion, CGH on cDNA microarrays revealed hundreds of novel genes whose overexpression is attributable to gene amplification. These genes may provide insights to the clonal evolution and progression of breast cancer and highlight promising therapeutic targets.

## INTRODUCTION

Gene expression patterns revealed by cDNA microarrays have facilitated classification of cancers into biologically distinct categories, some of which may explain the clinical behavior of the tumors (1-6). Despite this progress in diagnostic classification, the molecular mechanisms underlying gene expression patterns in cancer have remained elusive, and the utility of gene expression profiling in the identification of specific therapeutic targets remains limited.

Accumulation of genetic defects is thought to underlie the clonal evolution of cancer. Identification of the genes that mediate the effects of genetic changes may be important by highlighting transcripts that are actively involved in tumor progression. Such transcripts and their encoded proteins would be ideal targets for anticancer therapies, as demonstrated by the clinical success of new therapies against amplified oncogenes, such as *ERBB2* and *EGFR* (7, 8), in breast cancer and other solid tumors. Besides amplifications of known oncogenes, over

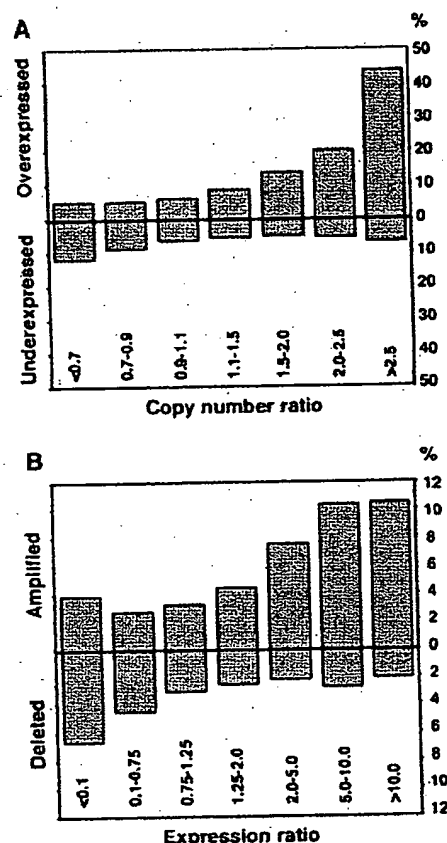


Fig. 1. Impact of gene copy number on global gene expression levels. *A*, percentage of over- and underexpressed genes (*Y* axis) according to copy number ratios (*X* axis). Threshold values used for over- and underexpression were  $>2.184$  (global upper 7% of the cDNA ratios) and  $<0.4826$  (global lower 7% of the expression ratios). *B*, percentage of amplified and deleted genes according to expression ratios. Threshold values for amplification and deletion were  $>1.5$  and  $<0.7$ .

20 recurrent regions of DNA amplification have been mapped in breast cancer by CGH<sup>5</sup> (9, 10). However, these amplicons are often large and poorly defined, and their impact on gene expression remains unknown.

We hypothesized that genome-wide identification of those gene expression changes that are attributable to underlying gene copy number alterations would highlight transcripts that are actively involved in the causation or maintenance of the malignant phenotype. To identify such transcripts, we applied a combination of cDNA and CGH microarrays to: (a) determine the global impact that gene copy number variation plays in breast cancer development and progression; and (b) identify and characterize those genes whose mRNA expres-

Received 5/29/02; accepted 8/28/02.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

<sup>1</sup> Supported in part by the Academy of Finland, Emil Aaltonen Foundation, the Finnish Cancer Society, the Pirkanmaa Cancer Society, the Pirkanmaa Cultural Foundation, the Finnish Breast Cancer Group, the Foundation for the Development of Laboratory Medicine, the Medical Research Fund of the Tampere University Hospital, the Foundation for Commercial and Technical Sciences, and the Swedish Research Council.

<sup>2</sup> Supplementary data for this article are available at Cancer Research Online (<http://cancerres.aacrjournals.org>).

<sup>3</sup> Contributed equally to this work.

<sup>4</sup> To whom requests for reprints should be addressed, at Laboratory of Cancer Genetics, Institute of Medical Technology, Lenkkelijankatu 6, FIN-33520 Tampere, Finland. Phone: 358-3247-4125; Fax: 358-3247-4168; E-mail: anne.kallioniemi@uta.fi.

<sup>5</sup> The abbreviations used are: CGH, comparative genomic hybridization; FISH, fluorescence in situ hybridization; RT-PCR, reverse transcription-PCR.

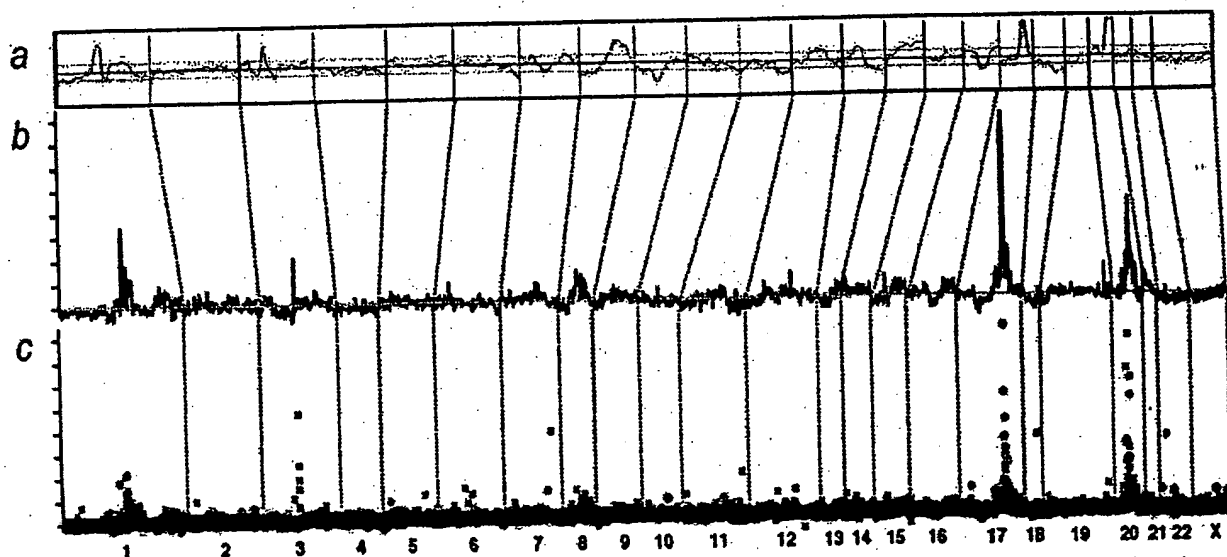


Fig. 2. Genome-wide copy number and expression analysis in the MCF-7 breast cancer cell line. *A*, chromosomal CGH analysis of MCF-7. The copy number ratio profile (blue line) across the entire genome from 1p telomere to Xq telomere is shown along with  $\pm 1$  SD (orange lines). The black horizontal line indicates a ratio of 1.0; red line, a ratio of 0.8; and green line, a ratio of 1.2. *B–C*, genome-wide copy number analysis in MCF-7 by CGH on cDNA microarray. The copy number ratios were plotted as a function of the position of the cDNA clones along the human genome. In *B*, individual data points are connected with a line, and a moving median of 10 adjacent clones is shown. Red horizontal line, the copy number ratio of 1.0. In *C*, individual data points are labeled by color coding according to cDNA expression ratios. The bright red dots indicate the upper 2%, and dark red dots, the next 5% of the expression ratios in MCF-7 cells (overexpressed genes); bright green dots indicate the lowest 2%, and dark green dots, the next 5% of the expression ratios (underexpressed genes); the rest of the observations are shown with black crosses. The chromosome numbers are shown at the bottom of the figure, and chromosome boundaries are indicated with a dashed line.

sion is most significantly associated with amplification of the corresponding genomic template.

## MATERIALS AND METHODS

**Breast Cancer Cell Lines.** Fourteen breast cancer cell lines (BT-20, BT-474, HCC1428, Hs578t, MCF7, MDA-361, MDA-436, MDA-453, MDA-468, SKBR-3, T-47D, UACC812, ZR-75-1, and ZR-75-30) were obtained from the American Type Culture Collection (Manassas, VA). Cells were grown under recommended culture conditions. Genomic DNA and mRNA were isolated using standard protocols.

**Copy Number and Expression Analyses by cDNA Microarrays.** The preparation and printing of the 13,824 cDNA clones on glass slides were performed as described (11–13). Of these clones, 244 represented uncharacterized expressed sequence tags, and the remainder corresponded to known genes. CGH experiments on cDNA microarrays were done as described (14, 15). Briefly, 20  $\mu$ g of genomic DNA from breast cancer cell lines and normal human WBCs were digested for 14–18 h with *Afl* and *Rsa* (Life Technologies, Inc., Rockville, MD) and purified by phenol/chloroform extraction. Six  $\mu$ g of digested cell line DNAs were labeled with Cy5-dUTP (Amersham Pharmacia) and normal DNA with Cy3-dUTP (Amersham Pharmacia) using the Bioprime Labeling kit (Life Technologies, Inc.). Hybridization (14, 15) and posthybridization washes (13) were done as described. For the expression analyses, a standard reference (Universal Human Reference RNA; Stratagene, La Jolla, CA) was used in all experiments. Forty  $\mu$ g of reference RNA were labeled with Cy3-dUTP and 3.5  $\mu$ g of test mRNA with Cy5-dUTP, and the labeled cDNAs were hybridized on microarrays as described (13, 15). For both microarray analyses, a laser confocal scanner (Agilent Technologies, Palo Alto, CA) was used to measure the fluorescence intensities at the target locations using the DEARRAY software (16). After background subtraction, average intensities at each clone in the test hybridization were divided by the average intensity of the corresponding clone in the control hybridization. For the copy number analysis, the ratios were normalized on the basis of the distribution of ratios of all targets on the array and for the expression analysis on the basis of 88 housekeeping genes, which were spotted four times onto the array. Low quality measurements (*i.e.*, copy number data with mean reference intensity <100 fluorescent units, and expression data with both test and reference intensity <100 fluorescent units and/or with spot size <50 units)

were excluded from the analysis and were treated as missing values. The distributions of fluorescence ratios were used to define cutpoints for increased/decreased copy number. Genes with CGH ratio >1.43 (representing the upper 5% of the CGH ratios across all experiments) were considered to be amplified, and genes with ratio <0.73 (representing the lower 5%) were considered to be deleted.

**Statistical Analysis of CGH and cDNA Microarray Data.** To evaluate the influence of copy number alterations on gene expression, we applied the following statistical approach. CGH and cDNA calibrated intensity ratios were log-transformed and normalized using median centering of the values in each cell line. Furthermore, cDNA ratios for each gene across all 14 cell lines were median centered. For each gene, the CGH data were represented by a vector that was labeled 1 for amplification (ratio, >1.43) and 0 for no amplification. Amplification was correlated with gene expression using the signal-to-noise statistics (1). We calculated a weight,  $w_g$ , for each gene as follows:

$$w_g = \frac{m_{g1} - m_{g0}}{\sigma_{g1} + \sigma_{g0}}$$

where  $m_{g1}$ ,  $\sigma_{g1}$ , and  $m_{g0}$ ,  $\sigma_{g0}$  denote the means and SDs for the expression levels for amplified and nonamplified cell lines, respectively. To assess the statistical significance of each weight, we performed 10,000 random permutations of the label vector. The probability that a gene had a larger or equal weight by random permutation than the original weight was denoted by  $\alpha$ . A low  $\alpha$  (<0.05) indicates a strong association between gene expression and amplification.

**Genomic Localization of cDNA Clones and Amplicon Mapping.** Each cDNA clone on the microarray was assigned to a Unigene cluster using the Unigene Build 141.<sup>6</sup> A database of genomic sequence alignment information for mRNA sequences was created from the August 2001 freeze of the University of California Santa Cruz's GoldenPath database.<sup>7</sup> The chromosome and bp positions for each cDNA clone were then retrieved by relating these data sets. Amplicons were defined as a CGH copy number ratio >2.0 in at least two adjacent clones in two or more cell lines or a CGH ratio >2.0 in at least three adjacent clones in a single cell line. The amplicon start and end positions were

<sup>6</sup> Internet address: [http://research.nhgri.nih.gov/microarray/downloadable\\_cdna.html](http://research.nhgri.nih.gov/microarray/downloadable_cdna.html).  
<sup>7</sup> Internet address: [www.genome.ucsc.edu](http://www.genome.ucsc.edu).

Table 1 Summary of independent amplicons in 14 breast cancer cell lines by CGH microarray

Location	Start (Mb)	End (Mb)	Size (Mb)
1p13	132.79	132.94	0.2
1q21	173.92	177.25	3.3
1q22	179.28	179.57	0.3
3p14	71.94	74.66	2.7
7p12.1-7p11.2	55.62	60.95	5.3
7q31	125.73	130.96	5.2
7q32	140.01	140.68	0.7
8q21.11-8q21.13	86.45	92.46	6.0
8q21.3	98.45	103.05	4.6
8q23.3-8q24.14	129.88	142.15	12.3
8q24.22	151.21	152.16	1.0
9p13	38.65	39.25	0.6
13q22-q31	77.15	81.38	4.2
16q22	86.70	87.62	0.9
17q11	29.30	30.85	1.6
17q12-q21.2	39.79	42.80	3.0
17q21.32-q21.33	52.47	55.80	3.3
17q22-q23.3	63.81	69.70	5.9
17q23.3-q24.3	69.93	74.99	5.1
19q13	40.63	41.40	0.8
20q11.22	34.59	35.85	1.3
20q13.12	44.00	45.62	1.6
20q13.12-q13.13	46.45	49.43	3.0
20q13.2-q13.32	51.32	59.12	7.8

extended to include neighboring nonamplified clones (ratio, <1.5). The amplicon size determination was partially dependent on local clone density.

**FISH.** Dual-color interphase FISH to breast cancer cell lines was done as described (17). Bacterial artificial chromosome clone RP11-361K8 was labeled with SpectrumOrange (Vysis, Downers Grove, IL), and SpectrumGreen-labeled probe for *EGFR* was obtained from Vysis. SpectrumGreen-labeled chromosome 7 and 17 centromere probes (Vysis) were used as a reference. A tissue microarray containing 612 formalin-fixed, paraffin-embedded primary breast cancers (17) was applied in FISH analyses as described (18). The use of these specimens was approved by the Ethics Committee of the University of Basel and by the NIH. Specimens containing a 2-fold or higher increase in the number of test probe signals, as compared with corresponding centromere signals, in at least 10% of the tumor cells were considered to be amplified. Survival analysis was performed using the Kaplan-Meier method and the log-rank test.

**RT-PCR.** The *HOXB7* expression level was determined relative to *GAPDH*. Reverse transcription and PCR amplification were performed using Access RT-PCR System (Promega Corp., Madison, WI) with 10 ng of mRNA as a template. *HOXB7* primers were 5'-GAGCAGAGGGACTCGGACTT-3' and 5'-GCGTCAGGTAGCGATTGTAG-3'.

## RESULTS

**Global Effect of Copy Number on Gene Expression.** 13,824 arrayed cDNA clones were applied for analysis of gene expression and gene copy number (CGH microarrays) in 14 breast cancer cell lines. The results illustrate a considerable influence of copy number on gene expression patterns. Up to 44% of the highly amplified transcripts (CGH ratio, >2.5) were overexpressed (i.e., belonged to the global upper 7% of expression ratios), compared with only 6% for genes with normal copy number levels (Fig. 1A). Conversely, 10.5% of the transcripts with high-level expression (cDNA ratio, >10) showed increased copy number (Fig. 1B). Low-level copy number increases and decreases were also associated with similar, although less dramatic, outcomes on gene expression (Fig. 1).

**Identification of Distinct Breast Cancer Amplicons.** Base-pair locations obtained for 11,994 cDNAs (86.8%) were used to plot copy number changes as a function of genomic position (Fig. 2, Supplement Fig. A). The average spacing of clones throughout the genome was 267 kb. This high-resolution mapping identified 24 independent breast cancer amplicons, spanning from 0.2 to 12 Mb of DNA (Table 1). Several amplification sites detected previously by chromosomal

CGH were validated, with 1q21, 17q12-q21.2, 17q22-q23, 20q13.1, and 20q13.2 regions being most commonly amplified. Furthermore, the boundaries of these amplicons were precisely delineated. In addition, novel amplicons were identified at 9p13 (38.65-39.25 Mb), and 17q21.3 (52.47-55.80 Mb).

**Direct Identification of Putative Amplification Target Genes.** The cDNA/CGH microarray technique enables the direct correlation of copy number and expression data on a gene-by-gene basis throughout the genome. We directly annotated high-resolution CGH plots with gene expression data using color coding. Fig. 2C shows that most of the amplified genes in the MCF-7 breast cancer cell line at 1p13, 17q22-q23, and 20q13 were highly overexpressed. A view of chromosome 7 in the MDA-468 cell line implicates *EGFR* as the most highly overexpressed and amplified gene at 7p11-p12 (Fig. 3A). In BT-474, the two known amplicons at 17q12 and 17q22-q23 contained numerous highly overexpressed genes (Fig. 3B). In addition, several genes, including the homeobox genes *HOXB2* and *HOXB7*, were highly amplified in a previously undescribed independent amplicon at 17q21.3. *HOXB7* was systematically amplified (as validated by FISH, Fig. 3B, inset) as well as overexpressed (as verified by RT-PCR, data not shown) in BT-474, UACC812, and ZR-75-30 cells. Furthermore, this novel

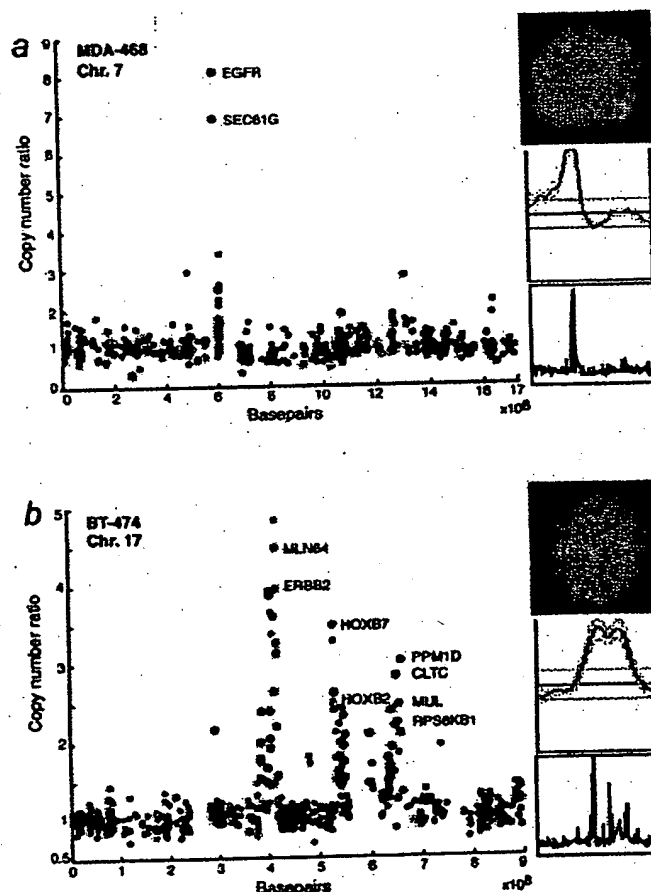
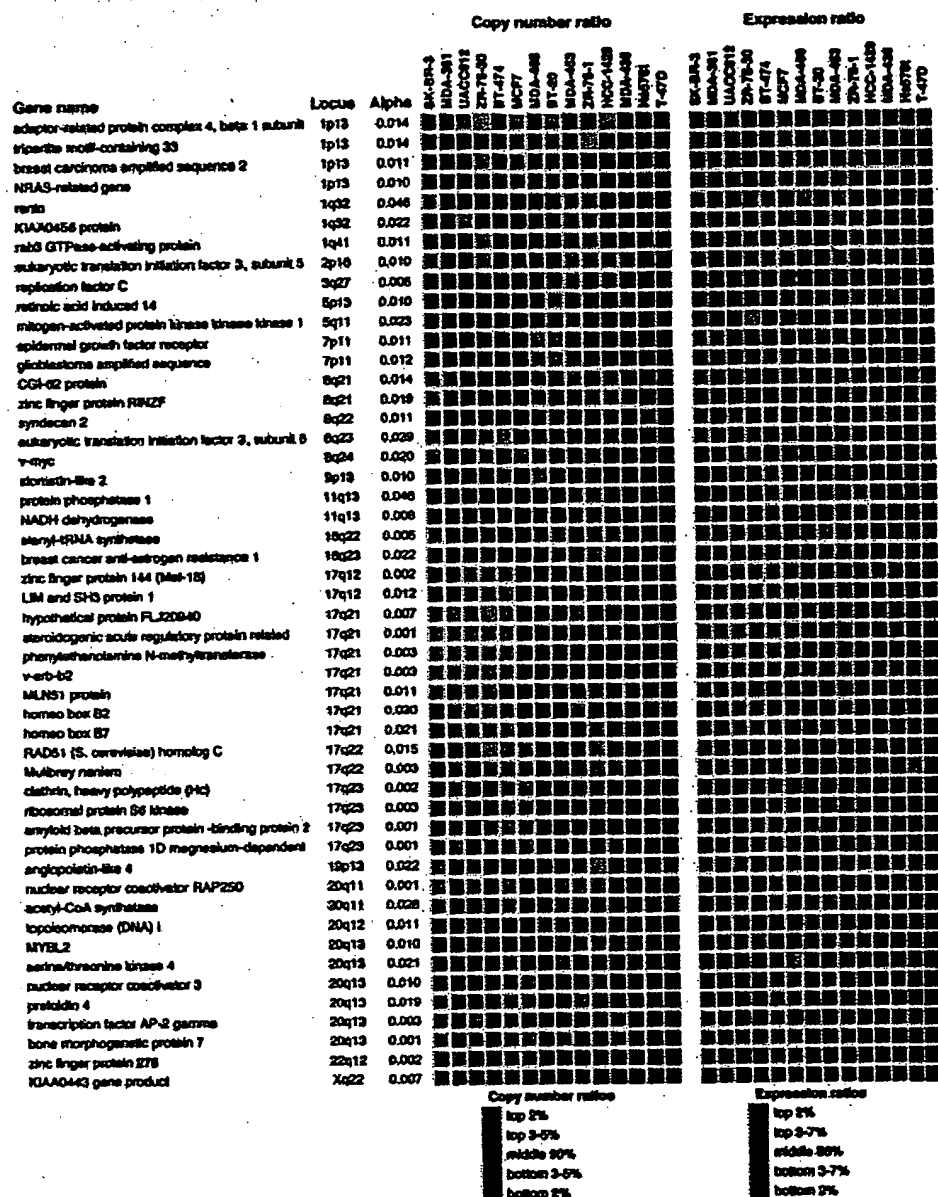


Fig. 3. Annotation of gene expression data on CGH microarray profiles. A, genes in the 7p11-p12 amplicon in the MDA-468 cell line are highly expressed (red dots) and include the *EGFR* oncogene. B, several genes in the 17q12, 17q21.3, and 17q23 amplicons in the BT-474 breast cancer cell line are highly overexpressed (red) and include the *HOXB7* gene. The data labels and color coding are as indicated for Fig. 2C. Insets show chromosomal CGH profiles for the corresponding chromosomes and validation of the increased copy number by interphase FISH using *EGFR* (red) and chromosome 7 centromere probe (green) to MDA-468 (A) and *HOXB7*-specific probe (red) and chromosome 17 centromere (green) to BT-474 cells (B).

Fig. 4. List of 50 genes with a statistically significant correlation ( $\alpha$  value  $<0.05$ ) between gene copy number and gene expression. Name, chromosomal location, and the  $\alpha$  value for each gene are indicated. The genes have been ordered according to their position in the genome. The color maps on the right illustrate the copy number and expression ratio patterns in the 14 cell lines. The key to the color code is shown at the bottom of the graph. Gray squares, missing values. The complete list of 270 genes is shown in supplemental Fig. B.



amplification was validated to be present in 10.2% of 363 primary breast cancers by FISH to a tissue microarray and was associated with poor prognosis of the patients ( $P = 0.001$ ).

**Statistical Identification and Characterization of 270 Highly Expressed Genes in Amplicons.** Statistical comparison of expression levels of all genes as a function of gene amplification identified 270 genes whose expression was significantly influenced by copy number across all 14 cell lines (Fig. 4, Supplemental Fig. B). According to the gene ontology data,<sup>8</sup> 91 of the 270 genes represented hypothetical proteins or genes with no functional annotation, whereas 179 had associated functional information available. Of these, 151 (84%) are implicated in apoptosis, cell proliferation, signal transduction, and transcription, whereas 28 (16%) had functional annotations that could not be directly linked with cancer.

## DISCUSSION

The importance of recurrent gene and chromosome copy number changes in the development and progression of solid tumors has been characterized in >1000 publications applying CGH<sup>9</sup> (9, 10), as well as in a large number of other molecular cytogenetic, cytogenetic, and molecular genetic studies. The effects of these somatic genetic changes on gene expression levels have remained largely unknown, although a few studies have explored gene expression changes occurring in specific amplicons (15, 19–21). Here, we applied genome-wide cDNA microarrays to identify transcripts whose expression changes were attributable to underlying gene copy number alterations in breast cancer.

The overall impact of copy number on gene expression patterns was substantial with the most dramatic effects seen in the case of high-

<sup>8</sup> Internet address: <http://www.geneontology.org/>.

<sup>9</sup> Internet address: <http://www.ncbi.nlm.nih.gov/entrez>.

level copy number increase. Low-level copy number gains and losses also had a significant influence on expression levels of genes in the regions affected, but these effects were more subtle on a gene-by-gene basis than those of high-level amplifications. However, the impact of low-level gains on the dysregulation of gene expression patterns in cancer may be equally important if not more important than that of high-level amplifications. Aneuploidy and low-level gains and losses of chromosomal arms represent the most common types of genetic alterations in breast and other cancers and, therefore, have an influence on many genes. Our results in breast cancer extend the recent studies on the impact of aneuploidy on global gene expression patterns in yeast cells, acute myeloid leukemia, and a prostate cancer model system (22–24).

The CGH microarray analysis identified 24 independent breast cancer amplicons. We defined the precise boundaries for many amplicons detected previously by chromosomal CGH (9, 10, 25, 26) and also discovered novel amplicons that had not been detected previously, presumably because of their small size (only 1–2 Mb) or close proximity to other larger amplicons. One of these novel amplicons involved the homeobox gene region at 17q21.3 and led to the overexpression of the *HOXB7* and *HOXB2* genes. The homeodomain transcription factors are known to be key regulators of embryonic development and have been occasionally reported to undergo aberrant expression in cancer (27, 28). *HOXB7* transfection induced cell proliferation in melanoma, breast, and ovarian cancer cells and increased tumorigenicity and angiogenesis in breast cancer (29–32). The present results imply that gene amplification may be a prominent mechanism for overexpressing *HOXB7* in breast cancer and suggest that *HOXB7* contributes to tumor progression and confers an aggressive disease phenotype in breast cancer. This view is supported by our finding of amplification of *HOXB7* in 10% of 363 primary breast cancers, as well as an association of amplification with poor prognosis of the patients.

We carried out a systematic search to identify genes whose expression levels across all 14 cell lines were attributable to amplification status. Statistical analysis revealed 270 such genes (representing ~2% of all genes on the array), including not only previously described amplified genes, such as *HER-2*, *MYC*, *EGFR*, ribosomal protein S6 kinase, and *AIB3*, but also numerous novel genes such as *NRAS-related gene* (1p13), *syndecan-2* (8q22), and *bone morphogenic protein* (20q13.1), whose activation by amplification may similarly promote breast cancer progression. Most of the 270 genes have not been implicated previously in breast cancer development and suggest novel pathogenetic mechanisms. Although we would not expect all of them to be causally involved, it is intriguing that 84% of the genes with associated functional information were implicated in apoptosis, cell proliferation, signal transduction, transcription, or other cellular processes that could directly imply a possible role in cancer progression. Therefore, a detailed characterization of these genes may provide biological insights to breast cancer progression and might lead to the development of novel therapeutic strategies.

In summary, we demonstrate application of cDNA microarrays to the analysis of both copy number and expression levels of over 12,000 transcripts throughout the breast cancer genome, roughly once every 267 kb. This analysis provided: (a) evidence of a prominent global influence of copy number changes on gene expression levels; (b) a high-resolution map of 24 independent amplicons in breast cancer; and (c) identification of a set of 270 genes, the overexpression of which was statistically attributable to gene amplification. Characterization of a novel amplicon at 17q21.3 implicated amplification and overexpression of the *HOXB7* gene in breast cancer, including a clinical association

between *HOXB7* amplification and poor patient prognosis. Overall, our results illustrate how the identification of genes activated by gene amplification provides a powerful approach to highlight genes with an important role in cancer as well as to prioritize and validate putative targets for therapy development.

## REFERENCES

1. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* (Wash. DC), 286: 531–537, 1999.
2. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* (Lond.), 403: 503–511, 2000.
3. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Sfezor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* (Lond.), 406: 536–540, 2000.
4. Perou, C. M., Sotlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Aksten, L. A., et al. Molecular portraits of human breast tumours. *Nature* (Lond.), 406: 747–752, 2000.
5. Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. Delineation of prognostic biomarkers in prostate cancer. *Nature* (Lond.), 412: 822–826, 2001.
6. Sotlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA*, 98: 10869–10874, 2001.
7. Ross, J. S., and Fletcher, J. A. The *HER-2/neu* oncogene: prognostic factor, predictive factor and target for therapy. *Semin. Cancer Biol.*, 9: 125–138, 1999.
8. Arteaga, C. L. The epidermal growth factor receptor: from mutant oncogene in nonhuman cancers to therapeutic target in human neoplasia. *J. Clin. Oncol.*, 19: 32–40, 2001.
9. Knuutila, S., Björkqvist, A. M., Autio, K., Tarkkanen, M., Wolf, M., Monni, O., Szymanska, J., Larramendy, M. L., Tapper, J., Pirc, H., El-Rifai, W., et al. DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. *Am. J. Pathol.*, 152: 1107–1123, 1998.
10. Knuutila, S., Autio, K., and Aalto, Y. Online access to CGH data of DNA sequence copy number changes. *Am. J. Pathol.*, 157: 689, 2000.
11. DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. Use of a cDNA microarray to analyze gene expression patterns in human cancer. *Nat. Genet.*, 14: 457–460, 1996.
12. Shalon, D., Smith, S. J., and Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.*, 6: 639–645, 1996.
13. Mousset, S., Bittner, M. L., Chen, Y., Dougherty, E. R., Baxevas, A., Meltzer, P. S., and Trent, J. M. Gene expression analysis by cDNA microarrays. In: F. J. Livesey and S. P. Hunt (eds.), *Functional Genomics*, pp. 113–137. Oxford: Oxford University Press, 2000.
14. Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, 23: 41–46, 1999.
15. Monni, O., Bärilund, M., Mousset, S., Kononen, J., Sauter, G., Heiskanen, M., Paavola, P., Avela, K., Chen, Y., Bittner, M. L., and Kallioniemi, A. Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. *Proc. Natl. Acad. Sci. USA*, 98: 5711–5716, 2001.
16. Chen, Y., Dougherty, E. R., and Bittner, M. L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, 2: 364–374, 1997.
17. Bärilund, M., Forozan, F., Kononen, J., Bubendorf, L., Chen, Y., Bittner, M. L., Thorst, J., Haas, P., Bucher, C., Sauter, G., et al. Detecting activation of ribosomal protein S6 kinase by complementary DNA and tissue microarray analysis. *J. Natl. Cancer Inst.*, 92: 1252–1259, 2000.
18. Andersen, C. L., Hostetter, G., Grigoryan, A., Sauter, G., and Kallioniemi, A. Improved procedure for fluorescence *in situ* hybridization on tissue microarrays. *Cytometry*, 45: 83–86, 2001.
19. Kauraniemi, P., Bärilund, M., Monni, O., and Kallioniemi, A. New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. *Cancer Res.*, 61: 8235–8240, 2001.
20. Clark, J., Edwards, S., John, M., Flohr, P., Gordon, T., Maillard, K., Giddings, I., Brown, C., Bagherzadeh, A., Campbell, C., Shipley, J., Wooster, R., and Cooper, C. S. Identification of amplified and expressed genes in breast cancer by comparative hybridization onto microarrays of randomly selected cDNA clones. *Genes Chromosomes Cancer*, 34: 104–114, 2002.
21. Varis, A., Wolf, M., Monni, O., Vakkari, M. L., Kakkola, A., Moskaluk, C., Frierson, H., Powell, S. M., Knuutila, S., Kallioniemi, A., and El-Rifai, W. Targets of gene amplification and overexpression at 17q in gastric cancer. *Cancer Res.*, 62: 2625–2629, 2002.
22. Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., Burchard, J., Dow, S., Ward, T. R., Kidcl, M. J., Friend, S. H., and Marton M. J.

- Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat. Genet.*, 25: 333-337, 2000.
23. Viraneva, K., Wright, F. A., Tanner, S. M., Yuan, B., Lemon, W. J., Caligiuri, M. A., Bloomfield, C. D., de La Chapelle, A., and Krahe, R. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl. Acad. Sci. USA*, 98: 1124-1129, 2001.
24. Phillips, J. L., Hayward, S. W., Wang, Y., Vasselli, J., Pavlovich, C., Padilla-Nash, H., Pezullo, J. R., Ghadimi, B. M., Grossfeld, G. D., Rivera, A., Linchan, W. M., Cunha, G. R., and Ried, T. The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. *Cancer Res.*, 61: 8143-8149, 2001.
25. Bärthel, M., Tirkkonen, M., Forozan, F., Tanner, M. M., Kallioniemi, O. P., and Kallioniemi, A. Increased copy number at 17q22-q24 by CGH in breast cancer is due to high-level amplification of two separate regions. *Genes Chromosomes Cancer*, 20: 372-376, 1997.
26. Tanner, M. M., Tirkkonen, M., Kallioniemi, A., Isola, J., Kuukasjärvi, T., Collins, C., Kowbel, D., Guan, X. Y., Trent, J., Gray, J. W., Meltzer, P., and Kallioniemi O. P. Independent amplification and frequent co-amplification of three nonsyntenic regions on the long arm of chromosome 20 in human breast cancer. *Cancer Res.*, 56: 3441-3445, 1996.
27. Cillo, C., Faiella, A., Cantile, M., and Boncinelli, E. Homeobox genes and cancer. *Exp. Cell Res.*, 248: 1-9, 1999.
28. Cillo, C., Cantile, M., Faiella, A., and Boncinelli, E. Homeobox genes in normal and malignant cells. *J. Cell. Physiol.*, 188: 161-169, 2001.
29. Care, A., Silvani, A., Meccia, E., Mattia, G., Stoppacciaro, A., Parmiani, G., Peschle, C., and Colombo, M. P. HOXB7 constitutively activates basic fibroblast growth factor in melanomas. *Mol. Cell. Biol.*, 16: 4842-4851, 1996.
30. Care, A., Silvani, A., Meccia, E., Mattia, G., Peschle, C., and Colombo, M. P. Transduction of the SkBr3 breast carcinoma cell line with the HOXB7 gene induces bFGF expression, increases cell proliferation and reduces growth factor dependence. *Oncogene*, 16: 3285-3289, 1998.
31. Care, A., Felicetti, F., Meccia, E., Bottero, L., Parenza, M., Stoppacciaro, A., Peschle, C., and Colombo, M. P. HOXB7: a key factor for tumor-associated angiogenic switch. *Cancer Res.*, 61: 6532-6539, 2001.
32. Naora, H., Yang, Y. Q., Montz, F. J., Seidman, J. D., Kurman, R. J., and Roden, R. B. A serologically identified tumor antigen encoded by a homeobox gene promotes growth of ovarian epithelial cells. *Proc. Natl. Acad. Sci. USA*, 98: 4060-4065, 2001.

# Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors

Jonathan R. Pollack<sup>\*†‡</sup>, Therese Sørli<sup>§</sup>, Charles M. Perou<sup>¶</sup>, Christian A. Rees<sup>||</sup>, Stefanie S. Jeffrey<sup>††</sup>, Per E. Lønning<sup>‡‡</sup>, Robert Tibshirani<sup>§§</sup>, David Botstein<sup>¶¶</sup>, Anne-Lise Børresen-Dale<sup>§</sup>, and Patrick O. Brown<sup>\*†¶</sup>

Departments of <sup>\*</sup>Pathology, <sup>¶</sup>Genetics, <sup>††</sup>Surgery, <sup>§§</sup>Health Research and Policy, and <sup>¶¶</sup>Biochemistry, and <sup>†</sup>Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; <sup>§</sup>Department of Genetics, Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway; <sup>††</sup>Department of Medicine (Oncology), Haukeland University Hospital, N-5021 Bergen, Norway; and <sup>‡‡</sup>Department of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599

Contributed by Patrick O. Brown, August 6, 2002

Genomic DNA copy number alterations are key genetic events in the development and progression of human cancers. Here we report a genome-wide microarray comparative genomic hybridization (array CGH) analysis of DNA copy number variation in a series of primary human breast tumors. We have profiled DNA copy number alteration across 6,691 mapped human genes, in 44 predominantly advanced, primary breast tumors and 10 breast cancer cell lines. While the overall patterns of DNA amplification and deletion corroborate previous cytogenetic studies, the high-resolution (gene-by-gene) mapping of amplicon boundaries and the quantitative analysis of amplicon shape provide significant improvement in the localization of candidate oncogenes. Parallel microarray measurements of mRNA levels reveal the remarkable degree to which variation in gene copy number contributes to variation in gene expression in tumor cells. Specifically, we find that 62% of highly amplified genes show moderately or highly elevated expression, that DNA copy number influences gene expression across a wide range of DNA copy number alterations (deletion, low-, mid- and high-level amplification), that on average, a 2-fold change in DNA copy number is associated with a corresponding 1.5-fold change in mRNA levels, and that overall, at least 12% of all the variation in gene expression among the breast tumors is directly attributable to underlying variation in gene copy number. These findings provide evidence that widespread DNA copy number alteration can lead directly to global deregulation of gene expression, which may contribute to the development or progression of cancer.

Conventional cytogenetic techniques, including comparative genomic hybridization (CGH) (1), have led to the identification of a number of recurrent regions of DNA copy number alteration in breast cancer cell lines and tumors (2–4). While some of these regions contain known or candidate oncogenes [e.g., *FGFR1* (8p11), *MYC* (8q24), *CCND1* (11q13), *ERBB2* (17q12), and *ZNF217* (20q13)] and tumor suppressor genes [*RB1* (13q14) and *TP53* (17p13)], the relevant gene(s) within other regions (e.g., gain of 1q, 8q22, and 17q22–24, and loss of 8p) remain to be identified. A high-resolution genome-wide map, delineating the boundaries of DNA copy number alterations in tumors, should facilitate the localization and identification of oncogenes and tumor suppressor genes in breast cancer. In this study, we have created such a map, using array-based CGH (5–7) to profile DNA copy number alteration in a series of breast cancer cell lines and primary tumors.

An unresolved question is the extent to which the widespread DNA copy number changes that we and others have identified in breast tumors alter expression of genes within involved regions. Because we had measured mRNA levels in parallel in the same samples (8), using the same DNA microarrays, we had an opportunity to explore on a genomic scale the relationship between DNA copy number changes and gene expression. From

this analysis, we have identified a significant impact of widespread DNA copy number alteration on the transcriptional programs of breast tumors.

## Materials and Methods

**Tumors and Cell Lines.** Primary breast tumors were predominantly large (>3 cm), intermediate-grade, infiltrating ductal carcinomas, with more than 50% being lymph node positive. The fraction of tumor cells within specimens averaged at least 50%. Details of individual tumors have been published (8, 9), and are summarized in Table 1, which is published as supporting information on the PNAS web site, [www.pnas.org](http://www.pnas.org). Breast cancer cell lines were obtained from the American Type Culture Collection. Genomic DNA was isolated either using Qiagen genomic DNA columns, or by phenol/chloroform extraction followed by ethanol precipitation.

**DNA Labeling and Microarray Hybridizations.** Genomic DNA labeling and hybridizations were performed essentially as described in Pollack *et al.* (7), with slight modifications. Two micrograms of DNA was labeled in a total volume of 50 microliters and the volumes of all reagents were adjusted accordingly. “Test” DNA (from tumors and cell lines) was fluorescently labeled (Cy5) and hybridized to a human cDNA microarray containing 6,691 different mapped human genes (i.e., UniGene clusters). The “reference” (labeled with Cy3) for each hybridization was normal female leukocyte DNA from a single donor. The fabrication of cDNA microarrays and the labeling and hybridization of mRNA samples have been described (8).

**Data Analysis and Map Positions.** Hybridized arrays were scanned on a GenePix scanner (Axon Instruments, Foster City, CA), and fluorescence ratios (test/reference) calculated using SCANALYZE software (available at <http://rana.lbl.gov>). Fluorescence ratios were normalized for each array by setting the average log fluorescence ratio for all array elements equal to 0. Measurements with fluorescence intensities more than 20% above background were considered reliable. DNA copy number profiles that deviated significantly from background ratios measured in normal genomic DNA control hybridizations were interpreted as evidence of real DNA copy number alteration (see *Estimating Significance of Altered Fluorescence Ratios* in the supporting information). When indicated, DNA copy number profiles are displayed as a moving average (symmetric 5-nearest neighbors). Map positions for arrayed human cDNAs were assigned by

Abbreviation: CGH, comparative genomic hybridization.

<sup>†</sup>To whom reprint requests should be addressed at: Department of Pathology, Stanford University School of Medicine, CCSR Building, Room 3245A, 269 Campus Drive, Stanford, CA 94305-5176. E-mail: [pollack1@stanford.edu](mailto:pollack1@stanford.edu).

<sup>††</sup>Present address: Zymyx Inc., Hayward, CA 94545.



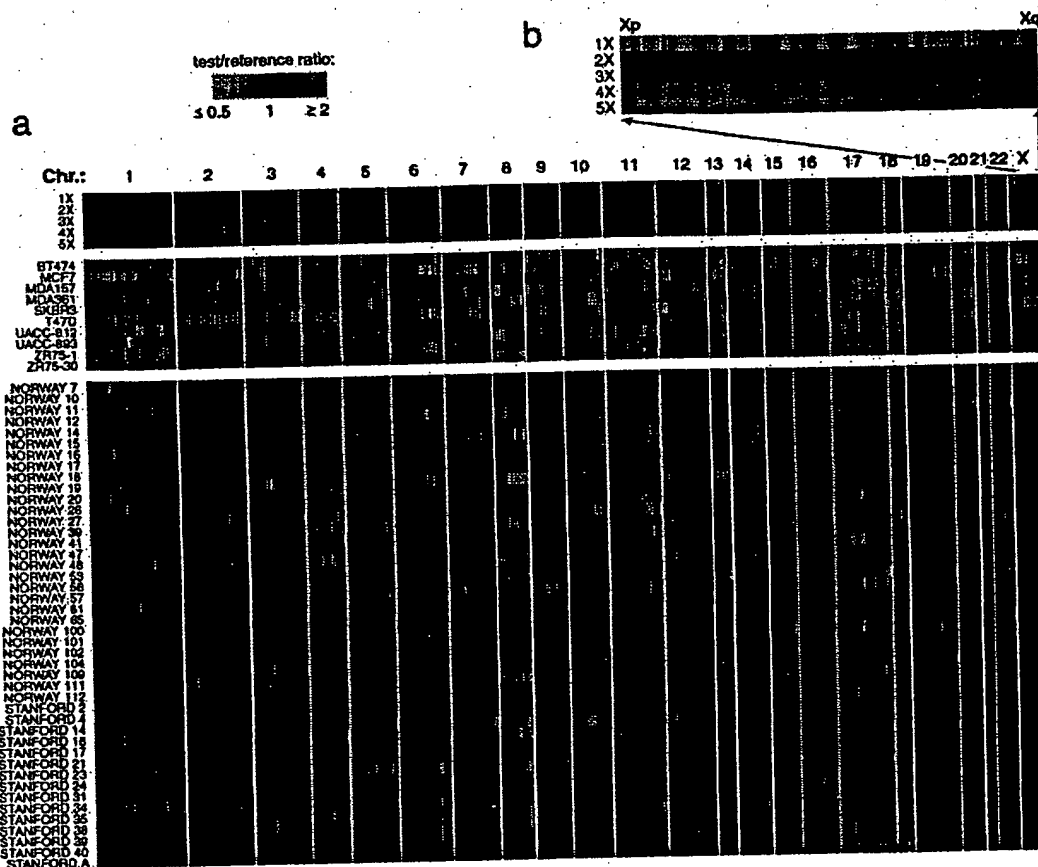


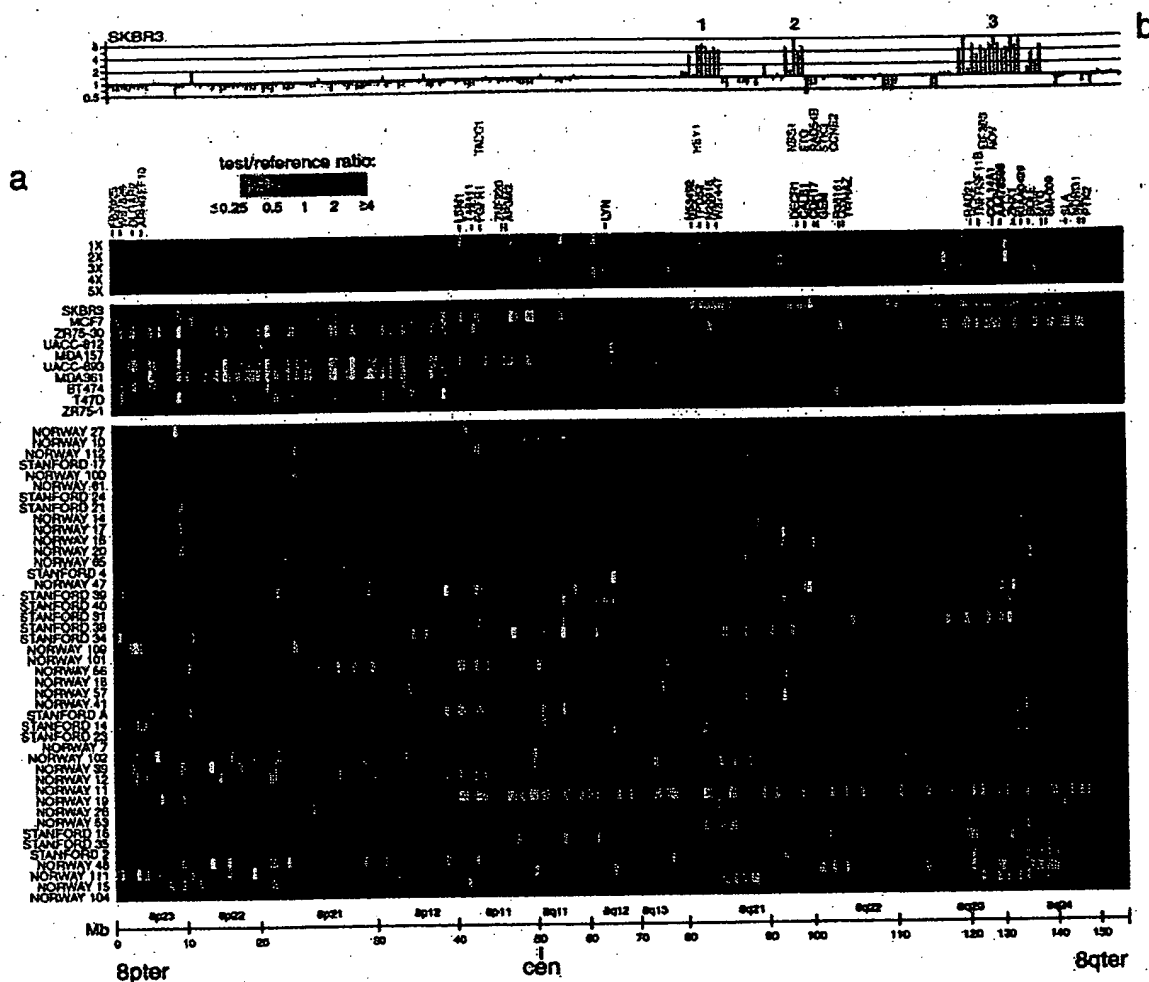
Fig. 1. Genome-wide measurement of DNA copy number alteration by array CGH. (a) DNA copy number profiles are illustrated for cell lines containing different numbers of X chromosomes, for breast cancer cell lines, and for breast tumors. Each row represents a different cell line or tumor, and each column represents one of 6,691 different mapped human genes present on the microarray, ordered by genome map position from 1pter through Xqter. Moving average (symmetric 5-nearest neighbors) fluorescence ratios (test/reference) are depicted using a log<sub>2</sub>-based pseudocolor scale (indicated), such that red luminescence reflects fold-amplification, green luminescence reflects fold-deletion, and black indicates no change (gray indicates poorly measured data). (b) Enlarged view of DNA copy number profiles across the X chromosome, shown for cell lines containing different numbers of X chromosomes.

identifying the starting position of the best and longest match of any DNA sequence represented in the corresponding UniGene cluster (10) against the "Golden Path" genome assembly (<http://genome.ucsc.edu/>; Oct 7, 2000 Freeze). For UniGene clusters represented by multiple arrayed elements, mean fluorescence ratios (for all elements representing the same UniGene cluster) are reported. For mRNA measurements, fluorescence ratios are "mean-centered" (i.e., reported relative to the mean ratio across the 44 tumor samples). The data set described here can be accessed in its entirety in the supporting information.

### Results

We performed CGH on 44 predominantly locally advanced, primary breast tumors and 10 breast cancer cell lines, using cDNA microarrays containing 6,691 different mapped human genes (Fig. 1a; also see *Materials and Methods* for details of microarray hybridizations). To take full advantage of the improved spatial resolution of array CGH, we ordered (fluorescence ratios for) the 6,691 cDNAs according to the "Golden Path" (<http://genome.ucsc.edu/>) genome assembly of the draft human genome sequences (11). In so doing, arrayed cDNAs not only themselves represent genes of potential interest (e.g., candidate oncogenes within amplicons), but also provide precise genetic landmarks for chromosomal regions of amplification and

deletion. Parallel analysis of DNA from cell lines containing different numbers of X chromosomes (Fig. 1b), as we did before (7), demonstrated the sensitivity of our method to detect single-copy loss (45, XO), and 1.5- (47,XXX), 2- (48,XXXX), or 2.5-fold (49,XXXXX) gains (also see Fig. 5, which is published as supporting information on the PNAS web site). Fluorescence ratios were linearly proportional to copy number ratios, which were slightly underestimated, in agreement with previous observations (7). Numerous DNA copy number alterations were evident in both the breast cancer cell lines and primary tumors (Fig. 1a), detected in the tumors despite the presence of euploid non-tumor cell types; the magnitudes of the observed changes were generally lower in the tumor samples. DNA copy-number alterations were found in every cancer cell line and tumor, and on every human chromosome in at least one sample. Recurrent regions of DNA copy number gain and loss were readily identifiable. For example, gains within 1q, 8q, 17q, and 20q were observed in a high proportion of breast cancer cell lines/tumors (90%/69%, 100%/47%, 100%/60%, and 90%/44%, respectively), as were losses within 1p, 3p, 8p, and 13q (80%/24%, 80%/22%, 80%/22%, and 70%/18%, respectively), consistent with published cytogenetic studies (refs. 2-4; a complete listing of gains/losses is provided in Tables 2 and 3, which are published as supporting information on the PNAS web site). The total



**Fig. 2.** DNA copy number alteration across chromosome 8 by array CGH. (a) DNA copy number profiles are illustrated for cell lines containing different numbers of X chromosomes, for breast cancer cell lines, and for breast tumors. Breast cancer cell lines and tumors are separately ordered by hierarchical clustering to highlight recurrent copy number changes. The 241 genes present on the microarrays and mapping to chromosome 8 are ordered by position along the chromosome. Fluorescence ratios (test/reference) are depicted by a log<sub>2</sub> pseudocolor scale (indicated). Selected genes are indicated with color-coded text (red, increased; green, decreased; black, no change; gray, not well measured) to reflect correspondingly altered mRNA levels (observed in the majority of the subset of samples displaying the DNA copy number change). The map positions for genes of interest that are not represented on the microarray are indicated in the row above those genes represented on the array. (b) Graphical display of DNA copy number profile for breast cancer cell line SKBR3. Fluorescence ratios (tumor/normal) are plotted on a log<sub>2</sub> scale for chromosome 8 genes, ordered along the chromosome.

number of genomic alterations (gains and losses) was found to be significantly higher in breast tumors that were high grade ( $P = 0.008$ ), consistent with published CGH data (3), estrogen receptor negative ( $P = 0.04$ ), and harboring TP53 mutations ( $P = 0.0006$ ) (see Table 4, which is published as supporting information on the PNAS web site).

The improved spatial resolution of our array CGH analysis is illustrated for chromosome 8, which displayed extensive DNA copy number alteration in our series. A detailed view of the variation in the copy number of 241 genes mapping to chromosome 8 revealed multiple regions of recurrent amplification; each of these potentially harbors a different known or previously uncharacterized oncogene (Fig. 2a). The complexity of amplicon structure is most easily appreciated in the breast cancer cell line SKBR3. Although a conventional CGH analysis of 8q in SKBR3 identified only two distinct regions of amplification (12), we observed three distinct regions of high-level amplification (labeled 1–3 in Fig. 2b). For each of these regions we can define the

boundaries of the interval recurrently amplified in the tumors we examined; in each case, known or plausible candidate oncogenes can be identified (a description of these regions, as well as the recurrently amplified regions on chromosomes 17 and 20, can be found in Figs. 6 and 7, which are published as supporting information on the PNAS web site).

For a subset of breast cancer cell lines and tumors (4 and 37, respectively), and a subset of arrayed genes (6,095), mRNA levels were quantitatively measured in parallel by using cDNA microarrays (8). The parallel assessment of mRNA levels is useful in the interpretation of DNA copy number changes. For example, the highly amplified genes that are also highly expressed are the strongest candidate oncogenes within an amplicon. Perhaps more significantly, our parallel analysis of DNA copy number changes and mRNA levels provides us the opportunity to assess the global impact of widespread DNA copy number alteration on gene expression in tumor cells.

A strong influence of DNA copy number on gene expression is evident in an examination of the pseudocolor representations

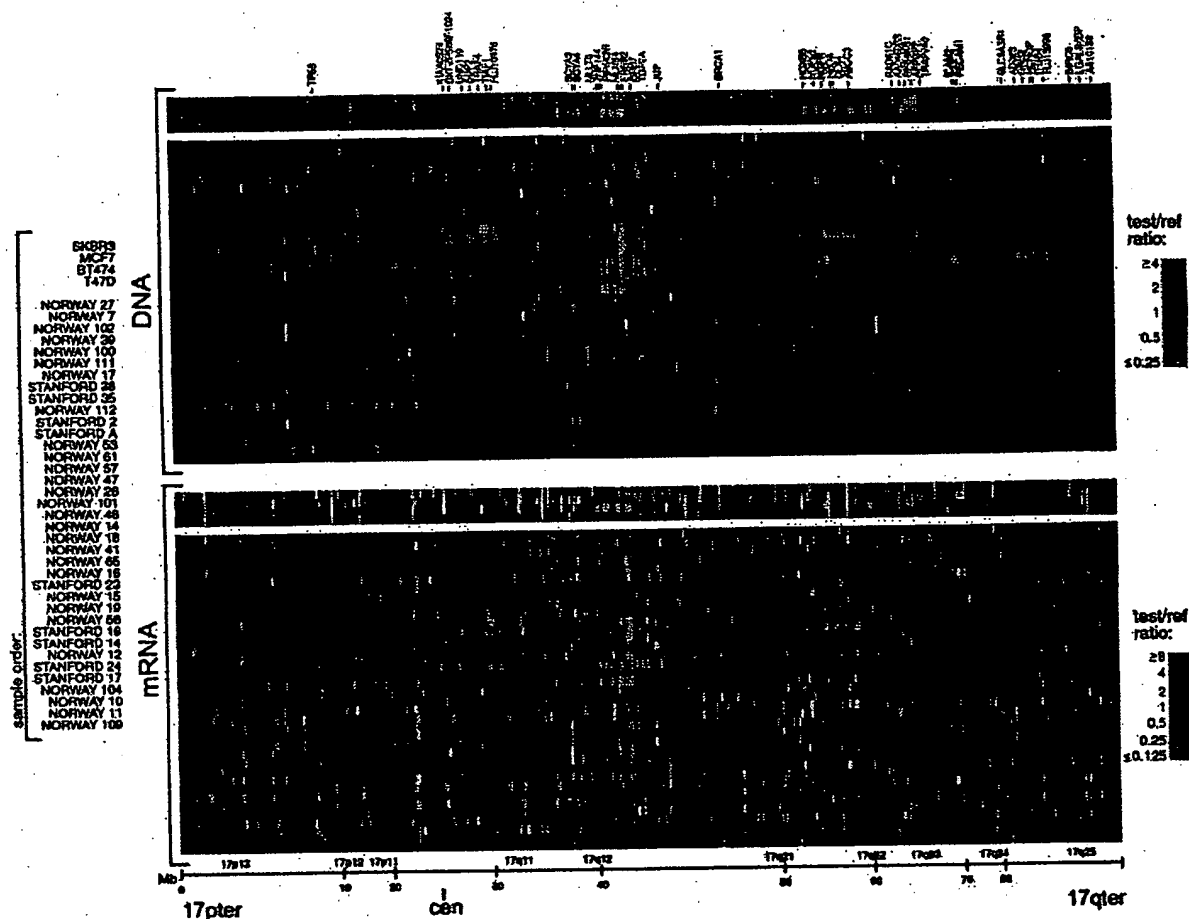


Fig. 3. Concordance between DNA copy number and gene expression across chromosome 17. DNA copy number alteration (Upper) and mRNA levels (Lower) are illustrated for breast cancer cell lines and tumors. Breast cancer cell lines and tumors are separately ordered by hierarchical clustering (Upper), and the identical sample order is maintained (Lower). The 354 genes present on the microarrays and mapping to chromosome 17, and for which both DNA copy number and mRNA levels were determined, are ordered by position along the chromosome; selected genes are indicated in color-coded text (see Fig. 2 legend). Fluorescence ratios (test/reference) are depicted by separate  $\log_2$  pseudocolor scales (indicated).

of DNA copy number and mRNA levels for genes on chromosome 17 (Fig. 3). The overall patterns of gene amplification and elevated gene expression are quite concordant; i.e., a significant fraction of highly amplified genes appear to be correspondingly highly expressed. The concordance between high-level amplification and increased gene expression is not restricted to chromosome 17. Genome-wide, of 117 high-level DNA amplifications (fluorescence ratios  $>4$ , and representing 91 different genes), 62% (representing 54 different genes; see Table 5, which is published as supporting information on the PNAS web site) are found associated with at least moderately elevated mRNA levels (mean-centered fluorescence ratios  $>2$ ), and 42% (representing 36 different genes) are found associated with comparably highly elevated mRNA levels (mean-centered fluorescence ratios  $>4$ ).

To determine the extent to which DNA deletion and lower-level amplification (in addition to high-level amplification) are also associated with corresponding alterations in mRNA levels, we performed three separate analyses on the complete data set (4 cell lines and 37 tumors, across 6,095 genes). First, we determined the average mRNA levels for each of five classes of genes, representing DNA deletion, no change, and low-, medium-, and high-level amplification (Fig. 4a). For both the

breast cancer cell lines and tumors, average mRNA levels tracked with DNA copy number across all five classes, in a statistically significant fashion ( $P$  values for pair-wise Student's  $t$  tests comparing adjacent classes: cell lines,  $4 \times 10^{-49}$ ,  $1 \times 10^{-49}$ ,  $5 \times 10^{-5}$ ,  $1 \times 10^{-2}$ ; tumors,  $1 \times 10^{-43}$ ,  $1 \times 10^{-214}$ ,  $5 \times 10^{-41}$ ,  $1 \times 10^{-4}$ ). A linear regression of the average  $\log(\text{DNA copy number})$ , for each class, against average  $\log(\text{mRNA level})$  demonstrated that on average, a 2-fold change in DNA copy number was accompanied by 1.4- and 1.5-fold changes in mRNA level for the breast cancer cell lines and tumors, respectively (Fig. 4a, regression line not shown). Second, we characterized the distribution of the 6,095 correlations between DNA copy number and mRNA level, each across the 37 tumor samples (Fig. 4b). The distribution of correlations forms a normal-shaped curve, but with the peak markedly shifted in the positive direction from zero. This shift is statistically significant, as evidenced in a plot of observed vs. expected correlations (Fig. 4c), and reflects a pervasive global influence of DNA copy number alterations on gene expression. Notably, the highest correlations between DNA copy number and mRNA level (the right tail of the distribution in Fig. 4b) comprise both amplified and deleted genes (data not shown). Third, we used a linear regression model to estimate the fraction of all variation measured in mRNA levels among the 37

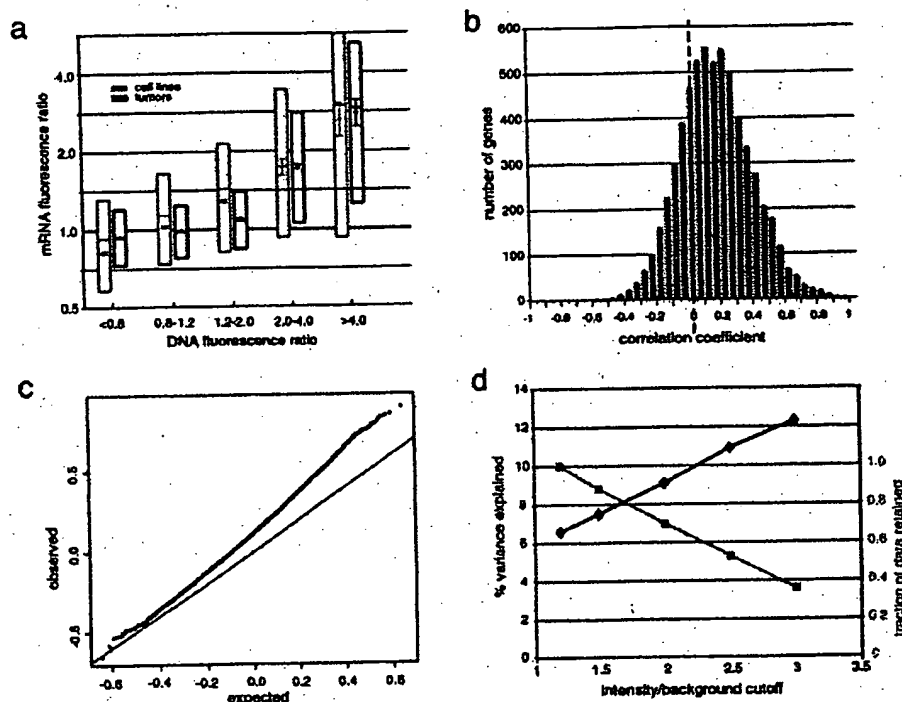


Fig. 4. Genome-wide influence of DNA copy number alterations on mRNA levels. (a) For breast cancer cell lines (gray) and tumor samples (black), both mean-centered mRNA fluorescence ratio (log<sub>2</sub> scale) quartiles (box plots indicate 25th, 50th, and 75th percentile) and averages (diamonds; Y-value error bars indicate standard errors of the mean) are plotted for each of five classes of genes, representing DNA deletion (tumor/normal ratio < 0.8), no change (0.8–1.2), low- (1.2–2), medium- (2–4), and high-level (>4) amplification. *P* values for pair-wise Student's *t* tests, comparing averages between adjacent classes (moving left to right), are  $4 \times 10^{-49}$ ,  $1 \times 10^{-49}$ ,  $5 \times 10^{-5}$ ,  $1 \times 10^{-2}$  (cell lines), and  $1 \times 10^{-43}$ ,  $1 \times 10^{-214}$ ,  $5 \times 10^{-41}$ ,  $1 \times 10^{-4}$  (tumors). (b) Distribution of correlations between DNA copy number and mRNA levels, for 6,095 different human genes across 37 breast tumor samples. (c) Plot of observed versus expected correlation coefficients. The expected values were obtained by randomization of the sample labels in the DNA copy number data set. The line of unity is indicated. (d) Percent variance in gene expression (among tumors) directly explained by variation in gene copy number. Percent variance explained (black line) and fraction of data retained (gray line) are plotted for different fluorescence intensity/background (a rough surrogate for signal/noise) cutoff values. Fraction of data retained is relative to the 1.2 intensity/background cutoff. Details of the linear regression model used to estimate the fraction of variation in gene expression attributable to underlying DNA copy number alteration can be found in the supporting information (see *Estimating the Fraction of Variation in Gene Expression Attributable to Underlying DNA Copy Number Alteration*).

tumors that could be attributed to underlying variation in DNA copy number. From this analysis, we estimate that, overall, about 7% of all of the observed variation in mRNA levels can be explained directly by variation in copy number of the altered genes (Fig. 4d). We can reduce the effects of experimental measurement error on this estimate by using only that fraction of the data most reliably measured (fluorescence intensity/background > 3); using that data, our estimate of the percent variation in mRNA levels directly attributed to variation in gene copy number increases to 12% (Fig. 4d). This still undoubtedly represents a significant underestimate, as the observed variation in global gene expression is affected not only by true variation in the expression programs of the tumor cells themselves, but also by the variable presence of non-tumor cell types within clinical samples.

#### Discussion

This genome-wide, array CGH analysis of DNA copy number alteration in a series of human breast tumors demonstrates the usefulness of defining amplicon boundaries at high resolution (gene-by-gene), and quantitatively measuring amplicon shape, to assist in locating and identifying candidate oncogenes. By analyzing mRNA levels in parallel, we have also discovered that changes in DNA copy number have a large, pervasive, direct effect on global gene expression patterns in both breast cancer

cell lines and tumors. Although the DNA microarrays used in our analysis may display a bias toward characterized and/or highly expressed genes, because we are examining such a large fraction of the genome (approximately 20% of all human genes), and because, as detailed above, we are likely underestimating the contribution of DNA copy number changes to altered gene expression, we believe our findings are likely to be generalizable (but would nevertheless still be remarkable if only applicable to this set of ~6,100 genes).

In budding yeast, aneuploidy has been shown to result in chromosome-wide gene expression biases (13). Two recent studies have begun to examine the global relationship between DNA copy number and gene expression in cancer cells. In agreement with our findings, Phillips *et al.* (14) have shown that with the acquisition of tumorigenicity in an immortalized prostate epithelial cell line, new chromosomal gains and losses resulted in a statistically significant respective increase and decrease in the average expression level of involved genes. In contrast, Platzer *et al.* (15) recently reported that in metastatic colon tumors only ~4% of genes within amplified regions were found more highly (>2-fold) expressed, when compared with normal colonic epithelium. This report differs substantially from our finding that 62% of highly amplified genes in breast cancer exhibit at least 2-fold increased expression. These contrasting findings may reflect methodological differences between the

studies. For example, the study of Platzer *et al.* (15) may have systematically under-measured gene expression changes. In this regard it is remarkable that only 14 transcripts of many thousand residing within unamplified chromosomal regions were found to exhibit at least 4-fold altered expression in metastatic colon cancer. Additionally, their reliance on lower-resolution chromosomal CGH may have resulted in poorly delimiting the boundaries of high-complexity amplicons, effectively overcalling regions with amplification. Alternatively, the contrasting findings for amplified genes may represent real biological differences between breast and metastatic colon tumors; resolution of this issue will require further studies.

Our finding that widespread DNA copy number alteration has a large, pervasive and direct effect on global gene expression patterns in breast cancer has several important implications. First, this finding supports a high degree of copy number-dependent gene expression in tumors. Second, it suggests that most genes are not subject to specific autoregulation or dosage compensation. Third, this finding cautions that elevated expression of an amplified gene cannot alone be considered strong independent evidence of a candidate oncogene's role in tumorigenesis. In our study, fully 62% of highly amplified genes demonstrated moderately or highly elevated expression. This highlights the importance of high-resolution mapping of amplicon boundaries and shape [to identify the "driving" gene(s) within amplicons (16)], on a large number of samples, in addition to functional studies. Fourth, this finding suggests that analyzing

the genomic distribution of expressed genes, even within existing microarray gene expression data sets, may permit the inference of DNA copy number aberration, particularly aneuploidy (where gene expression can be averaged across large chromosomal regions; see Fig. 3 and supporting information). Fifth, this finding implies that a substantial portion of the phenotypic uniqueness (and by extension, the heterogeneity in clinical behavior) among patients' tumors may be traceable to underlying variation in DNA copy number. Sixth, this finding supports a possible role for widespread DNA copy number alteration in tumorigenesis (17, 18), beyond the amplification of specific oncogenes and deletion of specific tumor suppressor genes. Widespread DNA copy number alteration, and the concomitant widespread imbalance in gene expression, might disrupt critical stoichiometric relationships in cell metabolism and physiology (e.g., proteasome, mitotic spindle), possibly promoting further chromosomal instability and directly contributing to tumor development or progression. Finally, our findings suggest the possibility of cancer therapies that exploit specific or global imbalances in gene expression in cancer.

We thank the many members of the P.O.B. and D.B. labs for helpful discussions. J.R.P. was a Howard Hughes Medical Institute Physician Postdoctoral Fellow during a portion of this work. P.O.B. is a Howard Hughes Medical Institute Associate Investigator. This work was supported by grants from the National Institutes of Health, the Howard Hughes Medical Institute, the Norwegian Cancer Society, and the Norwegian Research Council.

1. Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F. & Pinkel, D. (1992) *Science* 258, 818–821.
2. Kallioniemi, A., Kallioniemi, O. P., Piper, J., Tanner, M., Stokke, T., Chen, L., Smith, H. S., Pinkel, D., Gray, J. W. & Waldman, F. M. (1994) *Proc. Natl. Acad. Sci. USA* 91, 2156–2160.
3. Tirkkonen, M., Tanner, M., Karhu, R., Kallioniemi, A., Isola, J. & Kallioniemi, O. P. (1998) *Genes Chromosomes Cancer* 21, 177–184.
4. Forozan, F., Mahlamaki, E. H., Monni, O., Chen, Y., Veldman, R., Jiang, Y., Gooden, G. C., Ethier, S. P., Kallioniemi, A. & Kallioniemi, O. P. (2000) *Cancer Res.* 60, 4519–4525.
5. Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T. & Lichter, P. (1997) *Genes Chromosomes Cancer* 20, 399–407.
6. Pinkel, D., Segreaves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., *et al.* (1998) *Nat. Genet.* 20, 207–211.
7. Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D. & Brown, P. O. (1999) *Nat. Genet.* 23, 41–46.
8. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johansen, H., Akslen, L. A., *et al.* (2000) *Nature (London)* 406, 747–752.
9. Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* 98, 10869–10874.
10. Schuler, G. D. (1997) *J. Mol. Med.* 75, 694–698.
11. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature (London)* 409, 860–921.
12. Fejzo, M. S., Godfrey, T., Chen, C., Waldman, F. & Gray, J. W. (1998) *Genes Chromosomes Cancer* 22, 105–113.
13. Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., Burchard, J., Dow, S., Ward, T. R., Kidd, M. J., *et al.* (2000) *Nat. Genet.* 25, 333–337.
14. Phillips, J. L., Hayward, S. W., Wang, Y., Vasselli, J., Pavlovich, C., Padilla-Nash, H., Pezullo, J. R., Ghadimi, B. M., Grossfeld, G. D., Rivera, A., *et al.* (2001) *Cancer Res.* 61, 8143–8149.
15. Platzer, P., Upender, M. B., Wilson, K., Willis, J., Lutterbaugh, J., Nosrati, A., Willson, J. K., Mack, D., Ried, T. & Markowitz, S. (2002) *Cancer Res.* 62, 1134–1138.
16. Albertson, D. G., Ylstra, B., Segreaves, R., Collins, C., Dairkee, S. H., Kowbel, D., Kuo, W. L., Gray, J. W. & Pinkel, D. (2000) *Nat. Genet.* 25, 144–146.
17. Li, R., Yerganian, G., Duesberg, P., Kraemer, A., Willer, A., Rausch, C. & Hehlmann, R. (1997) *Proc. Natl. Acad. Sci. USA* 94, 14506–14511.
18. Rasnick, D. & Duesberg, P. H. (1999) *Biochem. J.* 340, 621–630.



# TECHNICAL UPDATE

FROM YOUR LABORATORY SERVICES PROVIDER

## HER-2/neu Breast Cancer Predictive Testing

*Julie Sanford Hanna, Ph.D. and Dan Mornin, M.D.*

EACH YEAR, OVER 182,000 WOMEN in the United States are diagnosed with breast cancer, and approximately 45,000 die of the disease.<sup>1</sup> Incidence appears to be increasing in the United States at a rate of roughly 2% per year. The reasons for the increase are unclear, but non-genetic risk factors appear to play a large role.<sup>2</sup>

Five-year survival rates range from approximately 65%-85%, depending on demographic group, with a significant percentage of women experiencing recurrence of their cancer within 10 years of diagnosis. One of the factors most predictive for recurrence once a diagnosis of breast cancer has been made is the number of axillary lymph nodes to which tumor has metastasized. Most node-positive women are given adjuvant therapy, which increases their survival. However, 20%-30% of patients without axillary node involvement also develop recurrent disease, and the difficulty lies in how to identify this high-risk subset of patients. These patients could benefit from increased surveillance, early intervention, and treatment.

Prognostic markers currently used in breast cancer recurrence prediction include tumor size, histological grade, steroid hormone receptor status, DNA ploidy, proliferative index, and cathepsin D status. Expression of growth factor receptors and over-expression of the HER-2/neu oncogene have also been identified as having value regarding treatment regimen and prognosis.

HER-2/neu (also known as c-erbB2) is an oncogene that encodes a transmembrane glycoprotein that is homologous to, but distinct from, the epidermal growth factor receptor. Numerous studies have indicated that high levels of expression of this protein are associated with rapid tumor growth, certain forms of therapy resistance, and shorter disease-free survival. The gene has been shown to be amplified and/or overexpressed in 10%-30% of invasive breast cancers and in 40%-60% of intraductal breast carcinoma.<sup>3</sup>

There are two distinct FDA-approved methods by which HER-2/neu status can be evaluated: immunohistochemistry (IHC, HercepTest™) and FISH (fluorescent in situ hybridization, PathVysion™ Kit). Both methods can be performed on archived and current specimens. The first method allows visual assessment of the amount of HER-2/neu protein present on the cell membrane. The latter method allows direct quantification of the level of gene amplification present in the tumor, enabling differentiation between low- versus high-amplification. At least one study has demonstrated a difference in

recurrence risk in women younger than 40 years of age for low- versus high-amplified tumors (54.5% compared to 85.7%); this is compared to a recurrence rate of 16.7% for patients with no HER-2/neu gene amplification.<sup>4</sup> HER-2/neu status may be particularly important to establish in women with small ( $\leq 1$  cm) tumor size.

The choice of methodology for determination of HER-2/neu status depends in part on the clinical setting. FDA approval for the Vysis FISH test was granted based on clinical trials involving 1549 node-positive patients. Patients received one of three different treatments consisting of different doses of cyclophosphamide, Adriamycin, and 5-fluorouracil (CAF). The study showed that patients with amplified HER-2/neu benefited from treatment with higher doses of adriamycin-based therapy, while those with normal HER-2/neu levels did not. The study therefore identified a sub-set of women, who because they did not benefit from more aggressive treatment, did not need to be exposed to the associated side effects. In addition, other evidence indicates that HER-2/neu amplification in node-negative patients can be used as an independent prognostic indicator for early recurrence, recurrent disease at any time and disease-related death.<sup>5</sup> Demonstration of HER-2/neu gene amplification by FISH has also been shown to be of value in predicting response to chemotherapy in stage-2 breast cancer patients.

Selection of patients for Herceptin® (Trastuzumab) monoclonal antibody therapy, however, is based upon demonstration of HER-2/neu protein overexpression using HercepTest™. Studies using Herceptin® in patients with metastatic breast cancer show an increase in time to disease progression, increased response rate to chemotherapeutic agents and a small increase in overall survival rate. The FISH assays have not yet been approved for this purpose, and studies looking at response to Herceptin® in patients with or without gene amplification status determined by FISH are in progress.

In general, FISH and IHC results correlate well. However, subsets of tumors are found which show discordant results; i.e., protein overexpression without gene amplification or lack of protein overexpression with gene amplification. The clinical significance of such results is unclear. Based on the above considerations, HER-2/neu testing at SHMC/PAML will utilize immunohistochemistry (HercepTest®) as a screen, followed by FISH in IHC-negative cases. Alternatively, either method may be ordered individually depending on the clinical setting or clinician preference.

## CPT code information

### HER-2/neu via IHC

88342 (including interpretive report)

### HER-2/neu via FISH

- 88271×2 Molecular cytogenetics, DNA probe, each  
88274 Molecular cytogenetics, interphase in situ hybridization, analyze 25-99 cells  
88291 Cytogenetics and molecular cytogenetics, interpretation and report

## Procedural Information

Immunohistochemistry is performed using the FDA-approved DAKO antibody kit, Herceptest®. The DAKO kit contains reagents required to complete a two-step immunohistochemical staining procedure for routinely processed, paraffin-embedded specimens. Following incubation with the primary rabbit antibody to human HER-2/neu protein, the kit employs a ready-to-use dextran-based visualization reagent. This reagent consists of both secondary goat anti-rabbit antibody molecules with horseradish peroxidase molecules linked to a common dextran polymer backbone, thus eliminating the need for sequential application of link antibody and peroxidase conjugated antibody. Enzymatic conversion of the subsequently added chromogen results in formation of visible reaction product at the antigen site. The specimen is then counterstained; a pathologist using light-microscopy interprets results.

FISH analysis at SHMC/PAML is performed using the FDA-approved PathVysion™ HER-2/neu DNA probe kit, produced by Vysis, Inc. Formalin fixed, paraffin-embedded breast tissue is processed using routine histological methods, and then slides are treated to allow hybridization of DNA probes to the nuclei present in the tissue section. The Pathvysion™ kit contains two direct-labeled DNA probes, one specific for the alphoid repetitive DNA (CEP 17, spectrum orange) present at the chromosome 17 centromere and the second for the HER-2/neu oncogene located at 17q11.2-12 (spectrum green). Enumeration of the probes allows a ratio of the number of copies of chromosome 17 to the number of copies of HER-2/neu to be obtained; this enables quantification of low versus high amplification levels, and allows an estimate of the percentage of cells with HER-2/neu gene amplification. The clinically relevant distinction is whether the gene amplification is due to increased gene copy number on the two chromosome 17 homologues normally present or an increase in the number of chromosome 17s in the cells. In the majority of cases, ratio equivalents less than 2.0 are indicative of a normal/negative result, ratios of 2.1 and over indicate that amplification is present and to what degree. Interpretation of this data will be performed and reported from the Vysis-certified Cytogenetics laboratory at SHMC.

## References

1. Wingo, P.A., Tong, T., Bolden, S., "Cancer Statistics", 1995;45:1:8-31.
2. "Cancer Rates and Risks", 4th ed., National Institutes of Health, National Cancer Institute, 1996, p. 120.
3. Slamon, D.J., Clark, G.M., Song, S.G., Levin, W.J., Ullrich, A., McGuire, W.L. "Human breast Cancer: Correlation of relapse and survival with amplification of the her-2/neu oncogene". Science, 235:177-182, 1987.
4. Xing, W.R., Gilchrist, K.W., Harris, C.P., Samson, W., Meisner, L.F. "FISH detection of HER-s/neu oncogene amplification in early onset breast cancer". Breast Cancer Res. And Treatment 39(2):203-212, 1996.
5. Press, M.F. Bernstein, L., Thomas, P.A., Meisner, L.F., Zhou, J.Y., Ma, Y., Hung, G., Robinson, R.A., Harris, C., El-Naggar, A., Slamon, D.J., Phillips, R.N., Ross, J.S., Wolman, S.R., Flom, K.J., "Her-2/neu gene amplification characterized by fluorescence in situ hybridization: poor prognosis in node-negative breast carcinomas", J. Clinical Oncology 15(8):2894-2904, 1997.

*Provided for the clients of*

**PATHOLOGY ASSOCIATES MEDICAL LABORATORIES  
PACLAB NETWORK LABORATORIES  
TRI-CITIES LABORATORY  
TREASURE VALLEY LABORATORY**

*For more information, please contact  
your local representative.*

## Genetic Instability in Epithelial Tissues at Risk for Cancer

WALTER N. HITTELMAN

Department of Experimental Therapeutics, The University of Texas  
M. D. Anderson Cancer Center, Houston, Texas 77030, USA

**ABSTRACT:** Epithelial tumors develop through a multistep process driven by genomic instability frequently associated with etiologic agents such as prolonged tobacco smoke exposure or human papilloma virus (HPV) infection. The purpose of the studies reported here was to examine the nature of genomic instability in epithelial tissues at cancer risk in order to identify tissue genetic biomarkers that might be used to assess an individual's cancer risk and response to chemopreventive intervention. As part of several chemoprevention trials, biopsies were obtained from risk tissues (i.e., bronchial biopsies from chronic smokers, oral or laryngeal biopsies from individuals with premalignancy) and examined for chromosome instability using *in situ* hybridization. Nearly all biopsy specimens show evidence for chromosome instability throughout the exposed tissue. Increased chromosome instability was observed with histologic progression in the normal to tumor transition of head and neck squamous cell carcinomas. Chromosome instability was also seen in premalignant head and neck lesions, and high levels were associated with subsequent tumor development. In bronchial biopsies of current smokers, the level of ongoing chromosome instability correlated with smoking intensity (e.g., packs/day), whereas the chromosome index (average number of chromosome copies per cell) correlated with cumulative tobacco exposure (i.e., pack-years). Spatial chromosome analyses of the epithelium demonstrated multifocal clonal outgrowths. In former smokers, random chromosome instability was reduced; however, clonal populations appeared to persist for many years, perhaps accounting for continued lung cancer risk following smoking cessation.

**KEYWORDS:** chromosome instability; epithelial cells; aerodigestive tract; chemoprevention; cancer risk

### THE NEED FOR BIOMARKERS OF CANCER RISK AND RESPONSE TO INTERVENTION

Epithelial cancers remain a major health challenge in the world. Despite improvements in staging and the application and integration of surgery, radiotherapy, and chemotherapy, the 5-year survival rate for individuals with lung cancer is only about 15%.<sup>1</sup> Even if strategies for early detection are successful and lung cancers are detected at a stage where local tumor resection and treatment is curative, these patients will still be at significant risk for developing second primary tumors

Address for correspondence: Dr. Walter N. Hittelman, Department of Experimental Therapeutics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd. (Box 19), Houston, Texas 77030. Voice: 713-792-2961; fax: 713-792-3754.  
whittelm@mdanderson.org



associated with the problem of field cancerization.<sup>2</sup> Similarly, for individuals with a first head and neck primary tumor, even if the first malignancy is successfully treated, the risk of developing a second primary in the tobacco smoke-exposed field is approximately 40%.<sup>3</sup> Similar cancer risk estimates exist for individuals who exhibit severe dysplasia in premalignant epithelial lesions.<sup>4</sup> For these reasons, it is important to focus on chemopreventive strategies to prevent the development of epithelial malignancies.

Several problems confront chemoprevention trials designed to identify efficacious agents.<sup>5</sup> First, chemoprevention trials with cancer incidence as a primary endpoint require tens of thousands of subjects and tens of years of intervention and follow-up for statistical evaluation. For example, a recently reported trial involved 30,000 subjects and required 10 years in order to examine the impact of prevention strategies on lung cancer development, only to find a possible increased lung cancer incidence in current smokers who received  $\beta$ -carotene.<sup>6</sup>

The problem of large, long-term trials results from the difficulty in identifying individuals at highest cancer risk who might best benefit from chemopreventive intervention. For example, 20 pack-year smokers, while known to be at relatively increased risk for developing lung cancer, have approximately a 10% lifetime risk for developing lung cancer.<sup>7</sup> This seriously limits the number of potentially useful strategies that can be clinically explored. A second problem facing chemoprevention trials is that little is known about what agents are likely to have efficacy, and even less is known regarding proper doses, schedules, and durations of treatment. Part of the reason for this problem is that too little is known about the physiologic processes that drive epithelial cancer development.

In order to reduce the number of subjects and the time required to carry out chemoprevention trials and thus allow the exploration of multiple prevention strategies, two types of advances are necessary. First, it is important to identify individuals at significantly increased cancer risk who might best benefit from different types of intervention. Second, in order to allow the rapid identification of agents, doses, and schedules of potentially efficacious agents, it is necessary to identify and validate surrogate endpoints of response that indicate whether the agents are having a positive impact on the target tissue during the chemopreventive intervention.

One approach to identifying individuals at increased aerodigestive tract cancer risk is to explore epidemiologic features of potential subjects. Molecular epidemiologic studies are beginning to identify intrinsic host factors that place some individuals at increased cancer risk, especially those with a chronic smoking history.<sup>8</sup> Most intrinsic factors identified thus far reflect levels of carcinogen metabolism, repair capabilities of the host following DNA damage, and other measures of intrinsic cellular sensitivity to mutagens. While these factors can provide statistically significant risk ratios in case-control studies that are controlled for tobacco exposure, the detected risk ratios usually fall in the range of 1.5 to 10. Unfortunately, this is not sufficient for the individualization of treatment and is not sufficiently high to significantly reduce the numbers of subjects required for chemoprevention trials with cancer incidence as the primary endpoint.

Another approach to identifying individuals at increased cancer risk is to directly examine the target tissue of individuals with known carcinogen exposure (e.g., chronic tobacco smoke exposure), who have evidence of target organ dysfunction

(e.g., chronic obstructive pulmonary disease, changes in voice quality), or who have clinical evidence of premalignancy (e.g., bronchial metaplasia/dysplasia, oral leukoplakia/erythroplakia, cervical intraepithelial neoplasia). The conventional standard for assessing cancer risk in these situations is the degree of histological change. However, while individuals who show moderate to severe dysplasia are known to be at increased cancer risk when compared to individuals with lesser histologic changes, it is often difficult to distinguish reactive changes to carcinogenic insult from initiated and progressing lesions. Similarly, upon cessation of carcinogenic insult, histologic changes may reverse yet cancer risk may continue for many years. For example, while smoking cessation is associated with decreased bronchial metaplasia,<sup>9</sup> increased lung cancer risk continues for many years beyond smoking cessation.<sup>10</sup> In fact, nearly half the newly diagnosed lung cancer cases in the USA occur in former smokers.<sup>11</sup>

The development of assays to identify individuals at high epithelial cancer risk and to directly assess response to intervention in the target tissue is therefore an important research goal. Such assays should be objective and easily quantifiable and, if possible, minimally invasive. Moreover, they should reflect both the disease process and the targeted pathway and thereby be useful in assessing risk and monitoring response to intervention as well as directly testing the hypothesized mechanism of action of the chemopreventive strategy.

In the chemoprevention setting it is important to recognize that one does not know the location of the future cancer. Thus, assays must necessarily be carried out on random biopsies of the field at risk. Even if there are clinically evident premalignant lesions, this does not mean that this is the likely site for a future malignancy. For example, nearly half of the cancers that develop in individuals with oral leukoplakia arise away from the original index lesion. Similarly, since many newly diagnosed lung cancers arise in the peripheral parts of the lung (e.g., adenocarcinomas), especially in former smokers, and since endobronchoscopy predominantly accesses central components of the lung, it is important to identify biomarkers that can reflect global processes ongoing in the target epithelial field associated with increased cancer risk. Their discovery requires a better understanding of the tumorigenesis process in epithelial fields at cancer risk.

#### THE RATIONALE FOR STUDYING GENOMIC INSTABILITY AS A MARKER OF RISK

Tumors of the aerodigestive tract have been proposed to reflect a "field cancerization" process whereby the whole tissue is exposed to carcinogenic insult (e.g., tobacco smoke) and is at increased risk for multistep tumor development.<sup>12,13</sup> Several types of clinical and laboratory data support this notion, including the frequent occurrence of synchronous primary and subsequent second primary tumors in the aerodigestive tract (frequently exhibiting dissimilar histologies as well as distinct genetic signatures<sup>14-16</sup>) and the presence of premalignant lesions that precede and/or accompany the tumor in the exposed tissue field.<sup>17</sup> The notion of a multistep tumorigenesis process is further supported by serial clinical and histologic evaluations of

target tissue or exfoliated cells where increasing degrees of histological abnormalities are observed over time.<sup>18</sup>

A working model for aerodigestive tract tumorigenesis is illustrated in FIGURE 1. Tumorigenesis in the face of carcinogenic exposure likely involves a chronic process of tissue injury and wound healing. DNA damage induced by the carcinogen is likely fixed into permanent genetic changes (e.g., chromosome damage, chromosome non-disjunction, gene mutation, gene deletion, etc.) during the process of proliferation. This damage would be expected to be distributed throughout the exposed tissue field leading to a background of generalized genomic damage (depicted in FIGURE 1 as a background mat of increasing density). Chronic injury and repair likely leads to the accumulation of cells with increasing amounts of genetic changes as well as the outgrowth of abnormal clones (triangles in FIGURE 1) carrying an accumulation of genetic changes important for selective survival, dysregulated growth, and preferential epithelial take-over by initiated clones (see FIGURE 2).

Cellular and molecular evidence for the field carcinogenesis and multistep tumorigenesis model comes from many laboratories.<sup>19,20</sup> With the advent of a wide array of molecular technologies, a large number of specific molecular genetic and epigenetic changes involving specific oncogenes, tumor suppressor genes, cell regulatory genes, and repair genes have now been described for aerodigestive tract cancers. The identification of these specific molecular changes have now provided probes to explore specific events occurring in premalignant lesions adjacent to aerodigestive tract tumors.<sup>21-24</sup> Frequently, these premalignant lesions showed a subset of the same molecular changes found in the associated tumor, suggesting that these lesions might represent precursor lesions for the associated tumors (i.e., a manifestation of

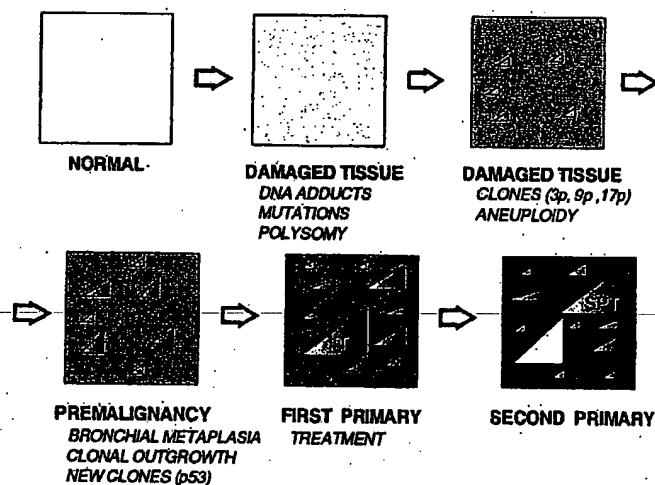


FIGURE 1. Field cancerization and multistep tumorigenesis.

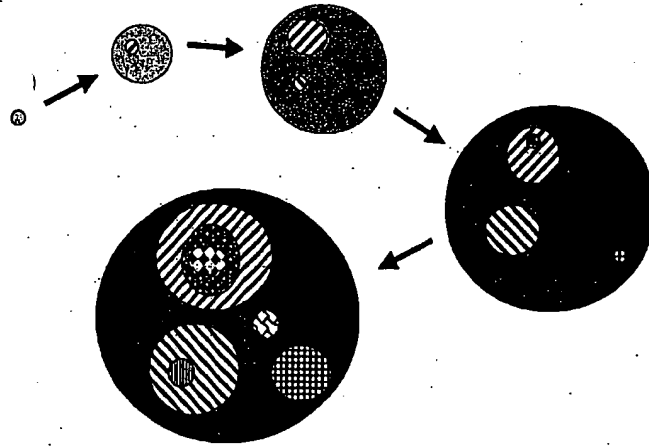


FIGURE 2. Multiple focal clonal evolution during multistep tumorigenesis.

a multistep tumorigenesis process). For example, studies of the premalignant lesions adjacent to head and neck tumors have provided evidence for a gradual accumulation of genetic alterations accompanied by evidence for dysregulation of cellular control mechanisms (e.g., alterations in expression of PCNA, EGFR, TGF- $\beta$ , p53, and cyclin D1).<sup>25-28</sup>

These types of studies have now also been applied to the target epithelium of individuals at increased risk for aerodigestive tract cancer (i.e., individuals with a chronic smoking/alcohol history and/or prior aerodigestive tract cancer). Several groups (using polymerase chain reaction, PCR, analysis of microdissected epithelium) have now demonstrated the presence of clonal outgrowths in the target premalignant epithelium of individuals at increased risk for cancer.<sup>29-31</sup> For example, examination of bronchial biopsies derived from individuals with a 20 pack-year smoking history demonstrated that 76% of the cases showed evidence for LOH (3p14, 9p21, or 17p13) in at least one of six lung biopsy sites. On a per site basis, some form of LOH was observed in 25% of the sites examined.<sup>29</sup>

If aerodigestive tract cancer development reflects a field cancerization process involving multistep events, then risk and response information should be able to be derived from random biopsies or exfoliated cells from the field at risk or from assessments of tissue undergoing similar processes. Hypothetically, lesions exhibiting the greatest degree of genomic instability, clonal outgrowth, and abnormal epithelial regulation would be at the highest relative aerodigestive tract cancer risk. Similarly, an active chemopreventive intervention might be expected to decrease these manifestations of risk. Reduced risk manifestations include decreased levels of ongoing genetic instability, decreased frequency of clonal outgrowths, and increased epithelial growth regulation.

### THE MEASUREMENT OF CHROMOSOME INSTABILITY USING CHROMOSOME *IN SITU* HYBRIDIZATION

Molecular genetic techniques, while extremely useful for detecting clonal changes in target tissues, are somewhat limited in their ability to detect random genetic instability. Conventional cytogenetic assays are useful for detecting chromosome instability and clonal chromosome changes. However, they require numbers of dividing cells for karyotypic analysis that are difficult to attain in the setting of biopsies acquired during the course of a chemoprevention trial. A technique was therefore needed that would allow chromosome instability measurements in situations where few cells are available (e.g. small biopsies, brushings, or sputum samples) and where the target material might be fixed. It was also desirable to have a technique that would be adaptable to tissue sections, whereby spatial information could be retained and genotype/phenotype associations could be determined on the same or adjacent tissue sections. The technique of *in situ* hybridization (ISH) involves the use of DNA probes that recognize either chromosome-specific repetitive target sequences, chromosome single gene copy sequences, or sequences along the whole chromosome length or chromosome segments.<sup>32</sup> We have adapted the ISH technique for formalin-fixed, paraffin-embedded tissue sections and have applied it to a variety of tissues, including the aerodigestive tract.<sup>33,34</sup>

Using probes that label the centromere regions of specific chromosomes, this assay permits determination of the average chromosome number per cell for each specimen. This assay is also useful for detecting generalized chromosome instability during the tumorigenesis process. Normal diploid populations should have two copies of each autosomal chromosome and should rarely show three or more chromosome copies per cell (chromosome polysomy), especially in tissue sections where nuclear truncation results in an under-representation of chromosome copy number. Thus, the detection of cells with three or more chromosome copies would indicate the presence of chromosome instability.

To examine this technique's potential for characterizing the multistep tumorigenesis process in the aerodigestive tract, we measured the fraction of cells exhibiting three or more chromosome copies in apparently contiguous epithelial transitions from normal to hyperplastic to dysplastic to carcinomas, all on a single tissue slice of head and neck squamous cell carcinomas.<sup>34</sup> In these specimens, greater than 35% of the cases of adjacent "normal" epithelium, greater than 65% of the cases of hyperplastic epithelium, and greater than 95% of the dysplastic and tumor regions showed evidence of chromosome polysomy. Of interest, similar transitions of chromosome instability were observed with at least four different chromosome probes. Similar trends have also been observed in amenable tissue from other epithelial malignancies, including cervix, bladder, and breast.<sup>35</sup> These results thus suggested that the notions of field cancerization and multistep tumorigenesis might apply to several epithelial tissues and that measures of chromosome instability might be useful for monitoring this process.

In the situations described above, the premalignant lesions examined might be considered to represent epithelium at 100% risk of being in a cancer field, since they were located in the adjacent epithelium to the cancer. This then raises the question of the nature of genetic instability in the epithelium of individuals at increased risk

for developing cancer. To explore this issue, we obtained biopsies during the course of leukoplakia chemoprevention trials exploring the use of 13-*cis*-retinoic acid in reversing leukoplakia and probed them for genetic instability using *in situ* hybridization. In one retrospective study and in one prospective study of subjects with oral leukoplakia, the results indicate that those subjects whose pretreatment biopsies harbor relatively high levels of genomic instability (i.e., more than 3% of the cells examined showing at least 3 chromosome 9 copies per cell) have a significantly higher likelihood of suffering early onset of head and neck cancer.<sup>36,37</sup> Interestingly, half of the tumors that did develop occurred away from the biopsy site used to measure genetic instability. This result suggests that genomic instability measurements in carcinogen-exposed tissue can provide useful cancer risk estimates.

#### THE RELATIONSHIP BETWEEN TOBACCO EXPOSURE AND CHROMOSOME INSTABILITY

In recent years, the aerodigestive tract chemoprevention group at M.D. Anderson Cancer Center has initiated three sequential biomarker-associated chemoprevention trials involving chronic smokers with a greater than 20 pack-year smoking history. In each of these studies, endobronchial biopsies were obtained from six defined sites within the lung, including the carina and at bifurcation points at the upper, middle, and lower right lung and at the upper and lower left lung. Biopsies were obtained prior to and following chemopreventive intervention and were subjected to *in situ* hybridization analysis in addition to analyses for other biomarkers. The first important finding was that some degree of chromosome polysomy was evident in all lung sites examined, and this was observed independently of the particular chromosome probe utilized.<sup>38</sup> This finding supports the notion that random chromosome changes may be occurring throughout the exposed lung field.

In a second study, bronchial biopsies were obtained from individuals with a 20 pack-year smoking history. In this study, most of the subjects involved were current smokers.<sup>39</sup> Interestingly, all cases who showed metaplasia at one of six biopsy sites also showed chromosome polysomy in at least one biopsy site; overall, 88% of the sites showed some evidence of chromosome 9 polysomy.<sup>40</sup> Evidence for genetic instability was also detected in patients who did not show evidence of bronchial metaplasia in any of six biopsy sites despite a strong smoking history. In fact, more than 90% of the cases and more than 60% of the sites showed significant chromosome polysomy (i.e., at least three copies in at least 2 % of the cells examined). These results suggest that the lungs of long-term smokers show significant evidence of genetic instability, and this instability can be detected throughout the accessible bronchial tree, even when bronchial metaplasia is not evident.

These studies in current smokers has allowed us to examine the relationship between the levels of genetic instability detected and subject characteristics such as smoking status (current or former), smoking history, and lung tissue pathologic changes. Evaluable biopsy material has now been obtained from more than 108 current smokers, including more than 480 evaluable biopsy sites. The mean metaplasia index in these current smokers was 30.4%. For the total population studied, the median chromosome index for the bronchial biopsies was 1.41 (range, 1.04–1.61).

and the median chromosome polysomy index was 2.0% (range 0–8.7%). This can be compared to a mean chromosome index between 1.2–1.4 for lymphocytes and very rare chromosome polysomy. Interestingly, the intrasubject variability in chromosome instability was relatively low in most subjects and was less than the intersubject variability. These results suggested that chronic smokers harbor detectable chromosome instability throughout the accessible bronchial tree (supporting the field carcinogenesis notion) and that information from one biopsy site might yield representative information for the rest of the lung field.

Since most of the current smokers exhibited bronchial metaplasia in at least one of the biopsied sites, this allowed us to examine the relationship between chromosome instability and histologic changes, both on a site-by-site basis and on a per case basis. On a site-by-site basis, the chromosome indices of lesions showing squamous metaplasia were similar to those not showing metaplasia (i.e., median 1.43 vs. 1.43), and the degree of chromosome polysomy in metaplastic lesions were only slightly higher than in non-metaplastic sites (medians: 2.2% vs. 1.8%, respectively). Thus, the presence or absence of squamous metaplasia at a biopsy site does not necessarily correlate with the degree of underlying genomic instability. On the other hand, those subjects with metaplasia indices of at least 15% also showed higher levels of chromosome polysomy than did subjects with metaplasia index below 15% (medians: 2.4% vs. 1.8%,  $p = 0.005$ ). Thus, these chromosome instability assessments in current smokers appeared to reflect a more global process in the lung field.

Tobacco exposure has been shown to significantly increase the risk of developing lung cancer, and the degree of risk is related to the extent of tobacco exposure. We were interested in determining the relationship between individuals' smoking history parameters and the levels of chromosome change found in their lungs following years of tobacco exposure. While there was significant intersubject variation for similar tobacco exposure histories, overall there was a significant correlation between the degree of chromosome polysomy and the intensity of ongoing tobacco exposure (packs/day,  $p = 0.02$  on a per site basis) and with the extent of tobacco exposure (pack-years,  $p = 0.003$ ). Thus the amount of chromosome polysomy reflects the intensity and extent of tobacco exposure. At the same time, individuals with similar smoking histories showed widely divergent amounts of chromosome polysomy, possibly reflecting differences in intrinsic sensitivity between subjects. There was also strong correlation between the chromosome index and the duration of the smoking history (smoking years) and total accumulated exposure (pack-years,  $p = 0.0001$ ). These results suggest that tobacco exposure is associated with the initiation and accumulation of chromosome instability in the exposed lung; however individuals are differentially sensitive to carcinogenic insult. The working hypothesis is that those individuals who accumulate the highest degree of chromosome changes will be at the highest lung cancer risk.

Many of the bronchial biopsies from chronic smokers examined by *in situ* hybridization showed a rise in the chromosome index above that expected for a diploid cell population, especially in subjects with an extensive smoking history. The rise in chromosome index was also accompanied by an increase in the fraction of cells exhibiting at least 3 chromosome copies per cell. To determine if a rise in the tissue chromosome index was due to clonal expansion of populations with chromosome trisomy, the chromosome copy number and relative coordinates of each cell scored in

the bronchial epithelium was recorded and a spatial genetic map was created.<sup>41</sup> We then developed algorithms for calculating localized chromosome indices within the tissue. Since trisomic clones would have, on average, three chromosomes instead of two, those cells involved in neighborhoods with chromosome indices three-halves that of diploid populations could be marked as being part of a trisomic clone. Similarly, groups of cells with chromosome indices half that of diploid populations could be marked as being part of a monosomic clone. This allowed the generation of a second-order, two-dimensional genetic map representation of the bronchial epithelium showing the relative locations of cells involved in monosomic and trisomic clonal outgrowths. When adjacent tissue sections from the same bronchial biopsy were probed separately for different chromosomes, the detected clones appeared to occupy separate subregions of the epithelium. This result suggests that not only are the lungs of chronic smokers undergoing a process of genetic instability, they are experiencing the outgrowth of multiple clones throughout the exposed lung field, as postulated by the models shown in FIGURES 1 and 2. One advantage of this clonal approach is that the contribution of both monosomic and multisomic clones can be detected.

Since smoking cessation has been suggested to reduce the lung cancer risk, it was of interest to determine whether the levels of chromosome instability would decrease following smoking cessation. This question was possible to examine because our third sequential chemoprevention trial involved subjects who had discontinued smoking. So far, more than 220 subjects (more than 650 biopsies) who have quit smoking (mean 9.9 quit-years) have been evaluated for chromosome instability in their lungs. Despite the fact that the mean metaplasia index in this group is 5.8% (considerably less than that in current smokers), chromosome instability is still observed in the majority of subjects.<sup>42</sup> While the mean chromosome polysomy level is reduced to 1.0%, some individuals continue to show polysomy levels above 5%. Interestingly, while the overall chromosome polysomy levels were reduced in these individuals who stopped smoking, the mean chromosome index remained at about 1.4 with some individuals exhibiting chromosome indices as high as 1.8. Initial chromosome mapping studies suggest that while random chromosome instability seems to decrease following smoking cessation, the clonal outgrowths may remain for many years in the lung. The working hypothesis is that those individuals who show the greatest degree of remaining chromosome instability are at the highest lung cancer risk despite smoking cessation. Long-term follow-up on these subjects will be necessary to test this hypothesis.

---

#### SUMMARY AND CONCLUSIONS

Aerodigestive tract tumorigenesis appears to be a multistep process taking place throughout the tissue fields of exposure. When viewed in the context of chromosome changes, carcinogen exposure appears to be associated with the random acquisition of chromosome polysomy throughout the exposed field, the degree of which is related to the degree and extent of carcinogen exposure as well as to the intrinsic susceptibility of the exposed individual. Continued exposure leads to continued acquisition of new changes and, in association with chronic wound-healing processes, to the



accumulation of clonal outgrowths throughout the target tissue. Although the ultimate malignancy may occur in only one or few tissue sites, manifestations of the instability process that drives tumorigenesis is globally present in the tissue. Thus random biopsies may provide useful risk information for the exposed field as a whole. Even when carcinogen exposure is reduced or chemopreventive strategies are initiated and histologic manifestations of the tumorigenesis process subside, the genetic scars of prior exposure remain in the form of clonal outgrowths and may explain continued lung cancer risk in ex-smokers. Future chemoprevention strategies need to focus on reducing the degree of chromosome instability and on trying to eliminate residual abnormal clonal outgrowths in the aerodigestive tract. In this setting, the measurement of chromosome instability in the target tissue will be useful in assessing cancer risk as well as response to intervention.

#### ACKNOWLEDGMENTS

The studies reviewed here represent one component of the collaborative efforts of the Aerodigestive Tract Chemoprevention team at The University of Texas M.D. Anderson Cancer Center, Houston, Texas. The studies were supported in part by National Institutes of Health-National Cancer Institute Grants CA-52051, CA-68437, CA 79437, CA 16672, CA 68089, CA 25433, CA 86390, CA 70907, NIH DE 13157, and the State of Texas Tobacco Research Fund.

#### REFERENCES

1. LANDIS, S.H., T. MURRAY, S. BOLDEN & P.A. WINGO. 1998. Cancer statistics, 1998. *CA Cancer J. Clin.* 48: 6-29.
2. JOHNSON, B.E. 1998. Second lung cancers in patients after treatment for an initial lung cancer. *J. Natl. Cancer Inst.* 90: 1335-1345.
3. LIPPMAN, S.M. & W.K. HONG. 1989. Second malignant tumors in head and neck squamous cell carcinoma: The overshadowing threat for patients with early stage of disease. *Int. J. Radiat. Oncol. Biol. Phys.* 17: 691-694.
4. SILVERMAN, S.J., JR., M. GORSKY & F. LOZADA. 1984. Oral leukoplakia and malignant transformation: a follow-up study of 257 patients. *Cancer* 53: 563-568.
5. LIPPMAN, S.M., J.S. LEE, R. LOTAN, *et al.* 1990. Biomarkers as intermediate endpoints in chemoprevention trials. *J. Natl. Cancer Inst.* 82: 555-560.
6. HEINONEN, O.P., D. ALBANES & THE ALPHA-TOCOPHEROL, BETA CAROTENE CANCER PREVENTION STUDY GROUP. 1994. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N. Engl. J. Med.* 330: 1029-1035.
7. PETO, R., S. DARBY, H. DEO, *et al.* 2000. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *Brit. Med. J.* 321: 323-329.
8. PERERA, F.P. 1996 Molecular epidemiology: insights into cancer susceptibility, risk assessment, and prevention. *J. Natl. Cancer Inst.* 88: 496-509.
9. LEE, J.S., S.M. LIPPMAN, S.E. BENNER, *et al.* 1994. Randomized placebo-controlled trial of isotretinoin in chemoprevention of bronchial squamous metaplasia. *J. Clin. Oncol.* 12: 937-941.

10. U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES. 1990. The health benefits of smoking cessation: a report of the Surgeon General. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. DHHS Pub. No. (CDC) 90-8416.
11. TONG, L., M.R. SPITZ, J.J. FAEGER, *et al.* 1996. Lung cancer in former smokers. *Cancer* 78: 1004-1010.
12. SLAUGHTER, D.P., H.W. SOUTHWICK & W. SMEJKAL. 1953. Field cancerization in oral stratified squamous epithelium: clinical implications of multicentric origin. *Cancer* 6: 963-968.
13. FARBER, E. 1984. The multistep nature of cancer development. *Cancer Res.* 44: 4217-4223.
14. CHUNG, K.Y., T. MUKHOPADHYAY, J. KIM, *et al.* 1993. Discordant p53 gene mutations in primary head and neck cancers and corresponding second primary cancers of the upper aerodigestive tract. *Cancer Res.* 53: 1676-1683.
15. SCHOLES, A.G.M., J.A. WOOLGAR, M.A. BOYLE, *et al.* 1998. Synchronous oral carcinomas: independent or common clonal origin? *Cancer Res.* 58: 2003-2006.
16. GLUCKMAN, J.O., J.D. CRISMAN & J.O. DONEGAN. 1980. Multicentric squamous cell carcinoma of the upper aerodigestive tract. *Head Neck Surg.* 3: 90-96.
17. AUERBACH, O., A.P. STOUT, E.C. HAMMOND, *et al.* 1961. Changes in bronchial epithelium in relation to cigarette smoking and in relation to lung cancer. *N. Engl. J. Med.* 265: 253-267.
18. SACCOMANNO, G., V.E. ARCHER, O. AUERBACH, *et al.* 1974. Development of carcinoma of the lung as reflected in exfoliated cells. *Cancer* 33: 256-270.
19. IZZO, J.G. & W.N. HITTELMAN. 1999. Characterization of multistep tumorigenesis by in situ hybridization. In *Introduction to Fluorescence In Situ Hybridization: Principles and Clinical Applications*. M. Andreff & D. Pinkel, Eds.: 173-208. John Wiley & Sons, Inc. New York.
20. HITTELMAN, W.N. 1999. Molecular cytogenetic evidence for multistep tumorigenesis: implications for risk assessment and early detection. In *Molecular Pathology of Cancer*. S. Srivastava, D.E. Hensen & A. Gazdar, Eds.: 385-404. IOS Press. Amsterdam, The Netherlands.
21. SUNDARESAN, V., P. GANLY, R. HASLETON, *et al.* 1992. p53 and chromosome 3 abnormalities, characteristic of malignant lung tumours, are detectable in preinvasive lesions of the bronchus. *Oncogene* 7: 1989-1997.
22. KISHIMOTO, Y., K. SUGIO, J.Y. HUNG, *et al.* 1995. Allele-specific loss in chromosome 9p loci in preneoplastic lesions accompanying non-small-cell lung cancers. *J. Natl. Cancer Inst.* 87: 1224-1229.
23. CALIFANO, J., P. VAN DER RIET, W. WESTRA, *et al.* 1996. Genetic progression model for head and neck cancer: implications for field cancerization. *Cancer Res.* 56: 2488-2492.
24. PARK I.W., I.I. WISTUBA, A. MAITRA, *et al.* 1999. Multiple clonal abnormalities in the bronchial epithelium of patients with lung cancer. *J. Natl. Cancer Inst.* 91: 1863-1868.
25. SHIN, D.M., N. VORAVUD, J.Y. RO, *et al.* 1994. Sequential increases in proliferating cell nuclear antigen expression in head and neck tumorigenesis: a potential biomarker. *J. Natl. Cancer Inst.* 85: 971-978.
26. SHIN, D.M., J.Y. RO, W.K. HONG, *et al.* 1994. Dysregulation of epidermal growth factor receptor expression in premalignant lesions during head and neck tumorigenesis. *Cancer Res.* 54: 3153-3159.
27. SHIN, D.M., J. KIM, J.Y. RO, *et al.* 1994. Activation of p53 gene expression in premalignant lesions during head and neck tumorigenesis. *Cancer Res.* 54: 321-326.
28. IZZO, J.G., V.A. PAPADIMITRAKOPOULOU, X.Q. LI, *et al.* 1998. Dysregulated cyclin D1 expression early in head and neck tumorigenesis: in vivo evidence for an association with subsequent gene amplification. *Oncogene* 17: 2313-2322.
29. MAO, L., J.S. LEE, J.M. KURIE, *et al.* 1997. Clonal genetic alterations in the lungs of current and former smokers. *J. Natl. Cancer Inst.* 89: 857-862.

30. WISTUBA, I.I., S. LAM, C. BEHRENS, *et al.* 1997. Molecular damage in the bronchial epithelium of current and former smokers. *J. Natl. Cancer Inst.* 89: 1366-1373.
31. MAO, L., J.S. LEE, Y.H. FAN, *et al.* 1996. Frequent microsatellite alterations at chromosomes 9p21 and 3p14 in oral premalignant lesions and their value in cancer risk assessment. *Nature Med.* 2: 682-685.
32. PODDIGHE, P.J., F.C. RAMAKERS & A.H. HOPMAN. 1992. Interphase cytogenetics of tumours. *J. Pathol.* 166: 215-224.
33. KIM, S.Y., J.S. LEE, J.Y. RO, *et al.* 1993. Interphase cytogenetics in paraffin sections of lung tumors by non-isotopic *in situ* hybridization. Mapping genotype/phenotype heterogeneity. *Am. J. Pathol.* 142: 307-317.
34. VORAVUD, N., D.M. SHIN, J.Y. RO, *et al.* 1993. Increased polysomies of chromosomes 7 and 17 during head and neck multistage tumorigenesis. *Cancer Res.* 53: 2874-2883.
35. HITTELMAN, W.N. 1999. Genetic instability assessments in the lung cancerization field. In *Lung Tumors: Fundamental Biology and Clinical Management*. C. Brambilla & E. Brambilla, Eds.: 255-267. Marcel Dekker. New York.
36. LEE, J.S., S.Y. KIM, W.K. HONG, *et al.* 1993. Detection of chromosomal polysomy in oral leukoplakia, a premalignant lesion. *J. Natl. Cancer Inst.* 85: 1951-1954.
37. LEE, J.J., W.K. HONG, W.N., HITTELMAN, *et al.* 2000. Predicting cancer development in oral leukoplakia: ten years of translational research. *Clin. Cancer Res.* 6: 1702-1710.
38. HITTELMAN W.N., R. YU, J. KURIE, *et al.* 1997. Evidence for genomic instability and clonal outgrowth in the bronchial epithelium of smokers [abstract]. *Proc. Am. Assoc. Cancer Res.* 38: 3097.
39. KURIE, J.M., J.S. LEE, F.R. KHURI, *et al.* N-(4-hydroxyphenyl)retinamide in the chemoprevention of squamous metaplasia and dysplasia of the bronchial epithelium. 2000. *Clin. Cancer Res.* 6: 2973-2979.
40. HITTELMAN, W.N., J.S. LEE, R.C. MORICE, *et al.* 1999. Lack of biomarker modulation in bronchial biopsies of chronic smokers following treatment with N-(4-hydroxyphenyl)retinamide (4-HPR). *Proc. Am. Assoc. Cancer Res.* 40: 2837.
41. HITTELMAN, W.N., J.S. LEE, N. CHEONG, *et al.* 1991. The chromosome view of "field cancerization" and multistep carcinogenesis. Implications for chemopreventive approaches. In *Chemoprevention of Cancer*. V. Pastorino & W.K. Hong, Eds.: 41-47. Georg Thieme Verlag. Stuttgart, Germany.
42. HITTELMAN, W.N., J.J. LEE, J.S. LEE, *et al.* 1998. Persistent genetic instability despite decreased proliferation in human lung tissue following smoking cessation. *Proc. AACR* 39: 336.

## Detection of Trisomy 7 in Nonmalignant Bronchial Epithelium from Lung Cancer Patients and Individuals at Risk for Lung Cancer<sup>1</sup>

Richard E. Crowell, Frank D. Gilliland, R. Thomas Temes, Heidi J. Harms, Robin E. Neft, Evelyn Heaphy, Dennis H. Auckley, Lida A. Crooks, Scott W. Jordan, Jonathan M. Samet, John F. Lechner, and Steven A. Belinsky<sup>2</sup>

Departments of Medicine [R. E. C., E. H., D. H. A.], Surgery [R. T. T.], and Pathology [L. A. C., S. W. J.], Albuquerque Veterans Administration Medical Center and the University of New Mexico Health Sciences Center, Albuquerque, New Mexico 87131; Inhalation Toxicology Research Institute, Albuquerque, New Mexico 87115 [H. J. H., R. E. N., J. F. L., S. A. B.]; Department of Epidemiology and Cancer Control Program, University of New Mexico Cancer Research and Treatment Center, Albuquerque, New Mexico 87131 [F. D. O.]; and Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland 21231 [J. M. S.]

### Abstract

Early identification and subsequent intervention are needed to decrease the high mortality rate associated with lung cancer. The examination of bronchial epithelium for genetic changes could be a valuable approach to identify individuals at greatest risk. The purpose of this investigation was to assay cells recovered from nonmalignant bronchial epithelium by fluorescence *in situ* hybridization for trisomy of chromosome 7, an alteration common in non-small cell lung cancer. Bronchial epithelium was collected during bronchoscopy from 16 cigarette smokers undergoing clinical evaluation for possible lung cancer and from seven individuals with a prior history of underground uranium mining. Normal bronchial epithelium was obtained from individuals without a prior history of smoking (never smokers). Bronchial cells were collected from a segmental bronchus in up to four different lung lobes for cytology and tissue culture. Twelve of 16 smokers were diagnosed with lung cancer. Cytological changes found in bronchial epithelium included squamous metaplasia, hyperplasia, and atypical glandular cells. These changes were present in 33, 12, and 47% of sites from lung cancer patients, smokers, and former uranium miners, respectively. Less than 10% of cells recovered from the diagnostic brush had cytological changes, and in several cases, these changes were present within different lobes from the same patient. Background

frequencies for trisomy 7 were  $1.4 \pm 0.3\%$  in bronchial epithelial cells from never smokers. Eighteen of 42 bronchial sites from lung cancer patients showed significantly elevated frequencies of trisomy 7 compared to never smoker controls. Six of the sites positive for trisomy 7 also contained cytological abnormalities. Trisomy 7 was found in six of seven patients diagnosed with squamous cell carcinoma, one of one patient with adenosquamous cell carcinoma, but in only one of four patients with adenocarcinoma. A significant increase in trisomy 7 frequency was detected in cytologically normal bronchial epithelium collected from four sites in one cancer-free smoker, whereas epithelium from the other smokers did not contain this chromosome abnormality. Finally, trisomy 7 was observed in almost half of the former uranium miners; three of seven sites positive for trisomy 7 also exhibited hyperplasia. Two of the former uranium miners who were positive for trisomy 7 developed squamous cell carcinoma 2 years after collection of bronchial cells. To determine whether the increased frequency of trisomy 7 reflects generalized aneuploidy or specific chromosomal duplication, a subgroup of samples was evaluated for trisomy of chromosome 2; the frequency was not elevated in any of the cases as compared with controls. The studies described in this report are the first to detect and quantify the presence of trisomy 7 in subjects at risk for lung cancer. These results also demonstrate the ability to detect genetic changes in cytologically normal cells, suggesting that molecular analyses may enhance the power for detecting premalignant changes in bronchial epithelium in high-risk individuals.

### Introduction

Although lung cancer is the leading cause of cancer death in the United States (1), early detection and intervention could decrease the high mortality rate associated with this disease if sensitive screening approaches could be developed (2-4). Early detection may be feasible because the entire respiratory tract is exposed to inhaled carcinogens; therefore, the whole lung is at risk for developing multiple, independently initiated sites. This "field cancerization" condition (5) is supported clinically by a high frequency of second primary tumors in lung cancer patients (6-9) and by the occurrence of progressive histological premalignant changes throughout the lower respiratory tract of cigarette smokers (10, 11). Moreover, recent studies using pathological tissues obtained after lung resection or autopsy have identified genetic aberrations associated with lung cancer in nonmalignant bronchial epithelium adjacent to tumors (12-16).

Although examination of pathological samples is useful for identifying genetic changes associated with carcinogenesis, this invasive approach for collection of clinical samples nec-

Received 1/23/96; revised 4/16/96; accepted 4/17/96.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

<sup>1</sup>This work was supported in whole or in part by the Office of Health and Environmental Research, United States Department of Energy, under Contracts DE-AC04-76V01013 and DE-FG03-92ER61520; by NIH Grant 5P50CA58184; and by the Dedicated Health Research Funds of the University of New Mexico School of Medicine.

<sup>2</sup>To whom requests for reprints should be addressed, at Inhalation Toxicology Research Institute, P. O. Box 5890, Albuquerque, NM 87185. Phone: (505) 845-1165; Fax: (505) 845-1198.

essayay for early detection would not be appropriate for screening. However, bronchial epithelial cells harvested using routine clinical procedures could be examined for genetic changes as an initial approach for detecting individuals at high risk for lung cancer. This approach could also provide genetic markers for evaluating the effectiveness of chemoprevention regimens. Bronchoscopy provides direct access to viable cells within the airways and is a commonly used tool for obtaining samples from the lower respiratory tract, including bronchial epithelium (17). This procedure can be used to repeatedly sample the bronchial epithelium over time and to collect viable cells that can be expanded through tissue culture for functional assays.

Because of field cancerization, genetic abnormalities should be dispersed throughout the bronchial epithelium of persons at risk for lung cancer. The purpose of this investigation was to test this hypothesis by sampling nonmalignant bronchial epithelium from distinct locations within four different lobes of the lung from persons at risk for lung cancer and then assaying the bronchial cells for the presence of specific genetic abnormalities. Trisomy of chromosome 7 was examined in these cells, because this alteration is common in solid tumors, including lung cancer, of several different organ systems (18, 19). In addition, trisomy 7 has been detected in premalignant lesions such as villous adenoma of the colon (20), in the colonic mucosa of individuals with familial polyposis (21), and in the far margins of some resected lung tumors (22). Our results demonstrate that trisomy 7 can be detected in nonmalignant bronchial epithelium from patients with lung cancer distant to the site of the tumor and in individuals without tumors who are at high risk for lung cancer development. Together, these studies suggest that an extra copy of chromosome 7 may be an intermediate biomarker of ongoing field carcinogenesis.

## Materials and Methods

**Subject Recruitment.** Bronchial epithelium was collected from 16 cigarette smokers undergoing a diagnostic workup for possible lung cancer and from 7 individuals with a prior history of underground uranium mining, 5 of whom were also smokers. Three individuals who had never smoked were also recruited to obtain bronchial epithelium not exposed directly to either tobacco smoke or radon progeny.

**Pathology and Exposure History.** Twelve of the 16 cigarette smokers who underwent diagnostic bronchoscopy were diagnosed with NSCLC.<sup>3</sup> Seven tumors were characterized histologically as SCCs, four tumors were ACs, and one tumor was an adenosquamous cell carcinoma. Lung cancer was not evident in the other four subjects. Smoking histories ranged from 15 to 120 pack-years (defined as the number of cigarettes smoked per day times the number of years smoked). All of the former uranium miners worked underground between 2 and 20 years, with a range of 27–527 working level months. Five of the seven miners had smoking histories that ranged from 20–60 pack-years.

**Bronchoscopic Collection and Processing of Bronchial Epithelium.** A protocol was developed for harvesting viable bronchial epithelium from the lower respiratory tract using a standard cytology brush during bronchoscopy. After introduc-

tion into the lower respiratory tract, the bronchoscope was directed into each upper and lower lobe, and the carinal margin of a segmental orifice, usually the second and third bifurcation within the upper and lower lobes, respectively, was brushed. These sites were chosen because (a) they are high-deposition areas for particles; (b) they are associated frequently with histological changes in smokers; and (c) they represent sites where tumors commonly occur (11, 23). The area was first washed with saline to remove any nonadherent cells. Sites were not brushed if a tumor was visualized within 5 cm of the site. After brushing, the brush was withdrawn, placed in serum-free medium, and kept on ice until processed. Each site was brushed twice. The procedure was well tolerated by all subjects, and no complications were noted related to the brushing procedure.

Bronchial cells were collected from only two of the sites in two of the subjects, from three sites in two subjects, and from all four sites in the remaining subjects. Although only two sites were brushed initially in case 1, cells were obtained from all four sites in this subject during a repeat bronchoscopy performed after the initial procedure did not yield a diagnosis. Samples were obtained from all four sites in the cancer-free current smokers and in the never smokers. In addition, bronchial epithelial cells derived at autopsy by Clonetics, Inc. (San Diego, CA) from four never smokers were also obtained to serve as additional controls. Only two sites sampled from most of the former uranium miners were available for analysis because cells recovered from the other sites had been used exclusively for cytology in another investigation.<sup>4</sup>

**Bronchial Epithelial Cell Culture.** Replicative cultures of the bronchial epithelial cells obtained by the procedure described above were established in our laboratory (24) using a serum-free medium (BEGM; Clonetics, Inc.) that is optimal for growth of these cells. Cells were removed from brushes by vigorous shaking in BEGM; cells from one brush were prepared for cytological analyses, and cells from the other brush were washed, resuspended in BEGM, seeded onto 60-mm fibronectin-coated plates, and grown at 37°C in 3% CO<sub>2</sub> and 21% O<sub>2</sub> until 80% confluence. Prior to passage, aliquots of cells were cryopreserved and stored at -145°C; other samples of cells were fixed in methanol-acetic acid (3:1). Next, the cells were washed four to six times in methanol:acetic acid and then dropped onto slides (about  $2 \times 10^5$  cells/slide). The effects of cell culture on the frequency of trisomy 7 in nonmalignant bronchial epithelium were examined by placing cells dispersed from brushes directly onto microscope slides followed by fixation.

**Cytology.** Cells from one brush from each bronchial collection site were prepared for cytological analysis by smearing the cells across a microscope slide. The cells were then fixed with 96% ethanol and stained according to the Papanicolaou procedure (25) to facilitate morphological evaluation by a cytopathologist.

**Detection of Trisomy 2 and Trisomy 7.** Trisomy 2 and trisomy 7 were determined by hybridization of cells with a biotinylated chromosome 2 or 7 centromere probe (Oncor, Gaithersburg, MD). The probes were denatured in hybridization buffer at 70°C for 5 min, and the slides were immersed in 70% formamide-2× SSPE at 70°C for 2 min. The probe was then applied to the slides, which were incubated in a humidified chamber at 37°C for 16 h. After incubation, the slides were washed in 0.25× SSPE (10 mM sodium phosphate monobasic monohydrate; 1 mM ethylenediamine tetraacetic acid disodium

**Data Analysis**  
nals in each  
frequency of  
the total num  
by the total r  
of the slides  
for trisomy 21  
positive for  
compared usi

## Results

**Cytology.** So far, the only cytological studies, were of the 10 patients, were covered from the 10 patients with cytological studies. The cytological abnormalities were detected from the 10 patients.

Two of the cytologically whereas no a three never : hyperplasia v four sites for people (Table 1). Culturing of establishing r chial brushing these cultures and does not fore, the cells potential was uranium min patients. Some passages (a n underwent 3C inhibited an inc

**Detection of**  
lium. Backg  
aminin norm  
from autopsy  
collected fro  
chial cell lin  
epithelial cell  
recruited nev  
cells containe  
with values r  
those reporte  
trisomy 7 fre  
controls) wen

Passage were examined (43%) samples for trisomy 7 at first passage (1). Three subjects (sites collected 7 and 12), trisomy 7. Six of the 18 samples were biologically abnormal.

<sup>3</sup> The abbreviations used are: NSCLC, non-small cell lung cancer; SCC, squamous cell cancer; AC, adenocarcinoma; EGFR, epidermal growth factor receptor; FISH, fluorescence *in situ* hybridization; LOH, loss of heterozygosity; BEGM, Bronchial Epithelium Growth Medium.

<sup>d</sup> Unpublished data.

salt, dihydrate; 150 mM sodium chloride, pH 7.4) for 5 min at 72°C, and the probe was detected with fluorescein-labeled avidin. Cell nuclei were visualized with propidium iodide.

**Data Analysis.** The number of centromeric hybridization signals in each cell were evaluated in 400 cells/slide, and the frequency of trisomy 7 on each slide was calculated by dividing the total number of cells expressing three hybridization signals by the total number of cells counted on each slide. Twenty % of the slides were scored by a second person, and frequencies for trisomy 7 differed by <0.4%. The total number of sites positive for trisomy 7 in subjects with SCC and AC were compared using Fisher's exact test.

## Results

**Cytology.** Squamous metaplasia and atypical glandular cells, the only cytological abnormalities observed in lung cancer patients, were present in 32% of the samples (Table 1). These cytological changes were observed in <10% of the cells recovered from the diagnostic brush. Two subjects had three sites with cytological abnormalities, and five subjects had no cytological abnormalities. No samples contained tumor cells by cytology, although one of four sites in five subjects was collected from the same lobe where a tumor was later diagnosed.

Two of the 16 sites in smokers without lung cancer were cytologically abnormal (both in the same person; Table 2), whereas no atypical cells were present in the 12 sites from the three never smokers (Table 3). In former uranium miners, hyperplasia was present in bronchial cells collected from all four sites from one person, and in one site in two additional people (Table 2).

**Culturing of Bronchial Epithelial Cells.** The efficiency of establishing replicative cultures of the cells obtained by bronchial brushing was 100%. The serum-free medium used for these cultures is optimal for growing bronchial epithelial cells and does not support fibroblastic cell replication (25). Therefore, the cells were uniformly epitheloid in appearance. Growth potential was evaluated by passaging cells from all seven of the uranium miner cases and cases 1-6 from the lung cancer patients. Some of these cultures were maintained for up to nine passages (a minimum of 16 population doublings), and many underwent 30 divisions before senescence. However, none exhibited an indefinite population-doubling potential.

**Detection of Trisomy 7 in Nonmalignant Bronchial Epithelium.** Background rates of trisomy 7 were determined by examining normal human bronchial epithelial cell lines derived from autopsy cases of never smokers and bronchial epithelium collected from never smokers during bronchoscopy. In bronchial cell lines (passage 2) from four donors and bronchial epithelial cell samples obtained by bronchial brushing from the recruited never smokers (Table 3), only  $1.4 \pm 0.3\%$  (SD) of the cells contained three hybridization signals for chromosome 7 with values ranging from 1 to 1.8%. These values agree with those reported by the manufacturer of the probe. Therefore, trisomy 7 frequencies of >2.0% (>2 SD above the mean for controls) were considered significantly different from controls.

Passage 1 or 2 bronchial cells from lung cancer patients were examined for trisomy 7. Eighteen of the 42 bronchial sites (43%) sampled from the 12 lung cancer patients contained trisomy 7 at frequencies ranging from 2.3 to 6.0% (Table 1; Fig. 1). Three subjects (cases 1, 2, and 11) displayed trisomy 7 in all sites collected during bronchoscopy, and in two subjects (cases 7 and 12), trisomy 7 was found in three of four sites (Table 1). Six of the 18 sites positive for trisomy 7 also contained cytologically abnormal cells. Trisomy 7 was found in six of seven

Table 1 Frequency of trisomy 7 in bronchial epithelial cells from lung cancer patients

Case	Age	Smoking (pack-yr)	Tumor diagnosis	Brush location	Cytological diagnosis	Trisomy 7 (frequency, %)
1	64	104	SCC	RLL <sup>a</sup>	N	2.8 <sup>b</sup>
				RUL	AGC	4.0 <sup>b</sup>
				RLL <sup>c</sup>	N	3.0 <sup>b</sup>
				RUL <sup>c</sup>	N	4.0 <sup>b</sup>
				LLL <sup>c</sup>	N	6.0 <sup>b</sup>
2	69	26	SCC	LLL <sup>c</sup>	SM	4.3 <sup>b</sup>
				RUL	SM	2.8 <sup>b</sup>
				LLL	SM	3.3 <sup>b</sup>
3	65	120	SCC	LUL	N	3.8 <sup>b</sup>
				RLL	AGC	2.0
				RUL	AGC	2.3 <sup>b</sup>
4	52	90	AC	LLL	AGC	2.0
				RLL	SM	1.5
				RUL	N	1.8
5	70	50	SCC	LLL	SM	1.5
				LUL	SM	1.8
				RLL	N	1.5
6	61	93	AC	RUL	N	1.5
				RUL	N	1.3
				LLL	N	2.0
7	58	40	SCC	LUL	N	1.5
				RLL	N	1.8
				RUL	N	2.3 <sup>b</sup>
8	59	120	AdSCC	LLL	N	2.5 <sup>b</sup>
				LUL	N	2.8 <sup>b</sup>
				RLL	N	1.5
9	65	71	SCC	RUL	N	2.0
				LLL	N	2.5 <sup>b</sup>
				LUL	AGC	2.0
10	63	45	AC	RLL	SM	2.0
				RUL	SM	2.5 <sup>b</sup>
				RLL	N	1.0
11	61	95	AC	RUL	N	1.8
				LLL	N	1.8
				LUL	N	1.3
12	76	17	SCC	LLL	N	2.5 <sup>b</sup>
				LUL	N	2.8 <sup>b</sup>
				RLL	N	2.0
				RUL	N	2.3 <sup>b</sup>
				LLL	N	2.3 <sup>b</sup>
				LUL	N	2.3 <sup>b</sup>
				LUL	N	2.3 <sup>b</sup>

<sup>a</sup> RLL, right lower lobe; RUL, right upper lobe; LLL, left lower lobe; LUL, left upper lobe; AGC, atypical glandular cells; SM, squamous metaplasia; N, normal cells; AdSCC, adenocarcinoma.

<sup>b</sup>  $P < 0.05$  as compared to never-smoker controls.

<sup>c</sup> Resampled 4 months later.

patients diagnosed with SCC, whereas only one of four patients with AC displayed trisomy 7 in any site collected at bronchoscopy. Case 7, which had histological features of both SCC and AC, had one site positive for trisomy 7. The frequency of positive trisomy 7 sites in all patients with SCC within this small sample population was significantly greater than in AC patients ( $P < 0.005$ ).

The reproducibility of detecting trisomy 7 at sites found to be positive for this abnormality was investigated in one patient (case 1) who required repeat bronchoscopy for clinical reasons. Trisomy 7 was increased similarly in the two sites brushed during both procedures, although cytological examination showed atypical cells in one site from the first bronchoscopy and cytologically normal cells from the same site collected

**Table 2** Frequency of trisomy 7 in bronchial epithelial cells from cancer-free smokers and former uranium miners

Case	Age	Smoking (pack-yr)	Radon exposure (WLMs) <sup>a</sup>	Brush location	Cytological diagnosis	Trisomy 7 (frequency, %)
13	81	15	0	RLL	N	1.8
				RUL	AGC	1.5
				LLL	N	1.8
				LUL	SM	2.0
14	34	24	0	RLL	N	1.3
				RUL	N	1.3
				LLL	N	1.0
				LUL	N	1.3
15	68	51	0	RLL	N	4.0 <sup>b</sup>
				RUL	N	3.0 <sup>b</sup>
				LLL	N	4.3 <sup>b</sup>
				LUL	N	3.5 <sup>b</sup>
16	45	30	0	RLL	N	1.3
				RUL	N	1.5
				LLL	N	2.0
				LUL	N	1.8
17	59	8	27	LLL	N	3.0 <sup>b</sup>
				LUL	N	3.0 <sup>b</sup>
18	65	9	516	LUL	N	1.3
				RUL	N	3.3 <sup>b</sup>
19	64	30	235	LUL	N	1.5
				RLL	N	1.0
20	56	0	186	LUL	N	2.0
				RLL	N	2.3 <sup>b</sup>
21	64	0	214	RLL	H	1.8
				LUL	N	1.8
22	64	9	577	RLL	H	0.8
				LLL	H	1.3
				LUL	H	2.8 <sup>b</sup>
				RLL	H	2.5 <sup>b</sup>
23	67	31	124	RUL	H	3.3 <sup>b</sup>

<sup>a</sup> Abbreviations are as indicated in Table 1 footnote. WLM, working level month; H, hyperplasia.

<sup>b</sup>  $P < 0.05$  as compared to never-smoker controls.

during the second procedure (Table 1). The other two sites collected during the second bronchoscopy also showed elevated frequencies of trisomy 7 in this patient.

Trisomy 7 was detected in cytologically normal bronchial epithelium collected from four sites in one (case 15) of the cancer-free smokers (Table 2). Bronchial cells from the other smokers did not contain this chromosome abnormality. In the former uranium miners (cases 17–23), seven of 15 sites collected during bronchoscopy were positive for trisomy 7. Three of the positive sites were found in one subject (case 23) and also contained basal cell hyperplasia. However, the other four samples positive for trisomy 7 showed no cytological abnormality.

Two of the former uranium miners (cases 18 and 23) developed lung cancer within 2 years of bronchial cell collection. SCC was diagnosed in the right upper lobe of both subjects. As noted in Table 2, both cases were positive for trisomy 7 in the right upper lobe brushing site obtained at the initial bronchoscopy.

**Tissue Culture Effects on Trisomy 7 Expression in Bronchial Epithelium.** The effect of tissue culture on trisomy 7 frequency was assessed by comparing the frequency of this chromosome abnormality in freshly isolated bronchial epithelium obtained directly from bronchial brushes ("preculture") to passage 1 cells. This comparison was conducted on cells collected from two different bronchial sites in three different subjects [(cases 11 and 16 and donor 7 (never smoker)]. Cultured samples positive for trisomy 7 in case 11 were also

**Table 3** Interphase analysis of chromosome 7 in normal human bronchial epithelial cells

Bronchial epithelial cell lines were established from never smokers (Clonetics) after autopsy and from volunteers. The normal distribution of chromosome 7 copy number as detected by FISH is shown by the percentage of cells exhibiting 1, 2, 3, or 4 hybridization signals. Four hundred cells containing hybridization signal were counted per donor.

Donor	Age	Brush location	Number of hybridization signals/cell (%)			
			1	2	3	4
1	6	NA <sup>a</sup>	3.5	92.0	1.5	3.0
		NA	2.3	95.5	1.3	1.0
2	17	NA	1.5	94.7	1.8	2.0
3	15	NA	2.0	94.8	1.0	2.3
4	41	NA	1.0	95.5	1.8	1.7
		RLL	0.5	98.3	1.0	0.2
5	45	RUL	1.3	96.5	1.0	1.2
		LLL	1.0	96.3	1.2	1.5
		LUL	1.0	96.8	1.0	1.2
		RLL	2.5	93.3	1.7	2.5
6	35	LLL	2.0	94.8	1.5	1.7
		LUL	1.8	94.2	1.8	2.2
		RLL	0.5	98.2	1.3	1.0
		RUL	1.2	96.8	1.3	0.7
7	33	LUL	1.0	96.0	1.5	1.5

<sup>a</sup> Abbreviations are as indicated in the legend to Table 1. NA, not applicable.

positive in preculture cells from the same bronchial collection site, whereas sites negative for trisomy 7 in cultured cells from case 16 and the never smoker were also negative in preculture cells (data not shown). Values for trisomy 7 differed by  $<0.3\%$  between preculture and cultured cells. The effect of passaging cells on the frequency of trisomy 7 was also examined in bronchial cells from case 1. Trisomy 7 frequency was similar in cells from passages 1, 4, and 7.

**Frequency of Trisomy 2 in Nonmalignant Bronchial Epithelium.** Aneuploidy has been detected in bronchial squamous metaplasia, a likely precursor to SCC (26). To determine whether the increased frequency of trisomy 7 detected in the current study reflects generalized aneuploidy or a specific chromosomal duplication, a subgroup of samples was evaluated for trisomy of chromosome 2. The frequency of trisomy 2 in never smokers was  $1.5 \pm 0.4\%$  (data not shown). Bronchial cells from eight subjects, six of whom had elevated frequencies for trisomy 7, were evaluated. The frequency for trisomy of chromosome 2 did not differ from never smokers (Table 4).

## Discussion

The studies described in this report are the first to detect and quantify an increase in trisomy 7 in the airway cells of subjects at risk for lung cancer. The presence of trisomy 7 appeared to be a specific chromosome gain and not due to generalized aneuploidy in these cells. In addition, trisomy 7 in nonmalignant epithelium from lung cancer patients was associated with SCC tumor histology, suggesting that patients with this genetic change may be at greater risk for developing SCC than other histological forms of lung cancer. This supposition was supported by the fact that two cancer-free former uranium miners with bronchial cells positive for trisomy 7 ultimately developed SCC. Finally, these results demonstrate the ability to detect genetic changes in cytologically normal cells, suggesting that molecular analyses may enhance the power for detecting

Fig. 1. FISH bronchial ep is apparent field. Magni

**Table 4** Fi

P

Case

1

2

7

8

13

15

19

23

Abbreviat

premalig

individu

Cig

ers to rac

cinogens

gens and

inhaled

Fig. 1. FISH for chromosome 7 in bronchial epithelial cells. Trisomy 7 is apparent in one cell from this field. Magnification,  $\times 530$ .

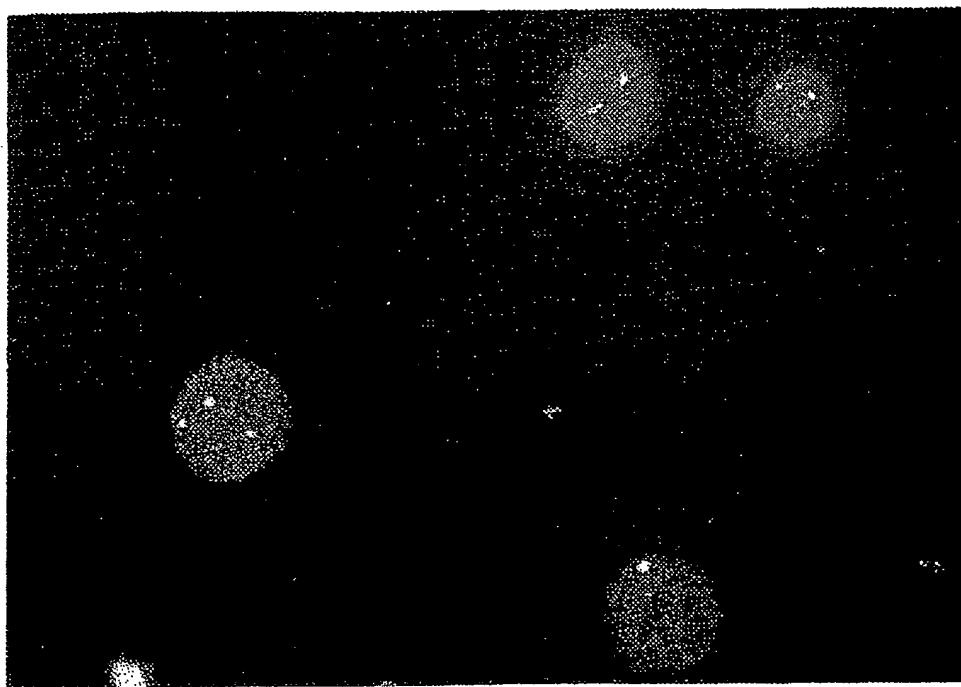


Table 4 Frequency of trisomy 2 in bronchial epithelial cells from lung cancer patients, cancer-free smokers, and former uranium miners

Case	Tumor diagnosis	Brush location	Trisomy 2 (frequency, %)
1	SCC	RLL*	1.5
		RUL	1.8
		LLL	1.8
		LUL	1.0
2	SCC	LLL	1.0
		LUL	1.0
		RLJ	1.5
7	SCC	LLL	1.8
		LUL	1.5
		RLL	0.3
8	AC	RUL	1.5
		LLL	0.8
		RLL	1.0
13	None	RUL	0.8
		LLL	1.0
		LUL	1.3
		RLL	1.8
15	None	RUL	2.0
		LLL	1.0
		LUL	1.3
		LUL	1.9
19	None	RLL	0.8
		RLJ	1.5
23	None	RLJ	1.5

\* Abbreviations are as indicated in legend to Table 1.

pre-malignant changes in bronchial epithelium in high-risk individuals.

Cigarette smoking and the exposure of underground miners to radon progeny are both well-established respiratory carcinogens (27, 28). Tobacco smoke contains numerous mutagens and carcinogens, and radon progeny that have been inhaled and deposited on the respiratory epithelium release  $\alpha$

particles capable of damaging DNA (28). Although comparison between findings in the cigarette smokers and the former uranium miners is constrained by the number of participants in the two groups, trisomy 7 was found in both groups. These results are consistent with the synergism between smoking and radon progeny, which suggests commonality in the pathways by which the two carcinogens cause lung cancer (29).

The bronchial brushing method used for collecting cells from the lower respiratory tract is rapid (10–12 min total for two brushes at four different sites), well tolerated by the patient, and permits collection of viable bronchial cells that can be expanded through tissue culture at 100% efficiency. The stability of these cells in culture was evident by the fact that the frequency of trisomy 7 did not differ between primary brush cells and cells propagated for up to seven passages. Furthermore, this procedure is amenable to the production of sufficient cell numbers ( $1 \times 10^8$ ) at low passage (one or two) to accommodate multiple molecular analyses. Although the media used in culturing of bronchial epithelial cells did not appear to provide a selective growth advantage to cells harboring an additional chromosome 7, the modulation of medium supplements might lead to the establishment of clonal populations of premalignant cells. Such cell populations would greatly facilitate the identification of additional early gene changes in respiratory carcinogenesis.

The detection of trisomy 7 in multiple nonmalignant sites within the bronchial tree supports the theory of field cancerization (5), which states that diffuse exposure of the entire respiratory tract to inhaled carcinogens causes the development of multiple, independently initiated sites that can lead to tumor development. Although the frequency of this chromosome abnormality was relatively low (2.3–6.0%), these values were consistent with the low percentage of cells within each brush sample (10%) that exhibited abnormal cytology. These results are also similar to studies of chromosome gain in patients with head and neck cancer where trisomy 7 was detected at frequen-



cies of 2, 3, and 21% in histologically normal, hyperplastic, and dysplastic cells, respectively (30).

The detection of trisomy 7 in normal, hyperplastic, and metaplastic bronchial epithelium from cancer-free patients extends a recent report describing LOH at chromosomes 3p, 5q, and 9p in dysplastic premalignant bronchial lesions harvested from current and former smokers by bronchoscopy (31). The inability to detect LOH at these chromosome loci in normal or early premalignant epithelium may stem from a difference in sensitivity between the methodologies used. The low frequency of trisomy 7 and cytologically abnormal cells collected from bronchoscopy is consistent with a lack of clonality within the brush cells. FISH assays on interphase cells permit screening of individual cells, and sensitivity for detection is limited only by the number of cells examined. In contrast, microsatellite analyses for LOH cannot detect nonclonal changes but require that the chromosome alteration be present in approximately 40–50% of the sample (32, 33).

The role of trisomy 7 in lung cancer development has not been elucidated. Increased expression of EGFR, which is located on chromosome 7 (34), is observed in 50–80% of NSCLCs (16, 35, 36). EGFR expression appears greater in SCC than AC (35, 36) and is amplified in some cell lines derived from SCC (37). These findings corroborate our hypothesis that acquisition of trisomy 7 in bronchial epithelium could be prognostic for development of SCC. Moreover, expression of this gene is also increased in nonmalignant bronchial epithelium from NSCLC patients (16, 35) and in normal or premalignant epithelium adjacent to head and neck tumors (38). Thus, altered expression of EGFR could enable cells that have acquired additional genetic changes to proliferate continually and escape from terminal differentiation (39). In addition, the *c-met* oncogene is also located on chromosome 7 and is overexpressed in NSCLCs (40, 41). This oncogene encodes a transmembrane tyrosine kinase (42) that functions as a receptor for the hepatocyte growth factor (43) and is involved in sustaining the growth of NSCLC cells in culture (44).

Previous studies have detected mutations in p53 (12, 14, 35), chromosome losses at 9p21 (45) and 3p (46) in preinvasive bronchial lesions, and simple chromosome rearrangements in normal bronchial epithelium from proximal airways (47) of lung cancer patients. The prevalence of these genetic changes in normal epithelium from persons at risk for lung cancer should be quantified by FISH to define the temporal sequences of somatic genetic changes that precede the development of clonal lesions in the lung. This information will be invaluable in providing biological markers that can qualitatively estimate the extent of field-cancerization in persons at risk for lung cancer and can be used to assess the efficacy of chemoprevention trials. Ultimately, the efficiency for detecting these biological markers in bronchial epithelium versus exfoliated epithelial cells within sputum must be established to support the use of a "genetic-based" screening approach for individuals at high risk for lung cancer. The results of the current investigation have identified one potential biomarker, trisomy 7, that may be useful in early detection and intervention for lung carcinogenesis.

## References

- Boring, C. C., Squires, T. S., and Tong, T. Cancer statistics, 1993. *J. Clin. Oncol.* 11: 1–26, 1993.
- Lippman, S. M., Benner, S. E., and Hong, W. K. Cancer chemoprevention. *J. Clin. Oncol.* 12: 851–873, 1994.
- Lippman, S. M., and Spitz, M. R. Intervention in the premalignant process. *Cancer Bull.* 43: 473–474, 1991.
- Berlin, N. I., Buncher, C. R., Fontana, R. S., Frost, J. K., and Melamed, M. R. The National Cancer Institute cooperative early lung cancer detection program. Early lung cancer detection. *Am. Rev. Respir. Dis.* 130: 545–570, 1984.
- Slaughter, D. P., Southwick, H. W., and Smejkal, W. Field cancerization in oral stratified squamous epithelium. Clinical implications of multicentric origin. *Cancer (Phila.)* 5: 963–968, 1953.
- van Horgeum, P. C., Wagenaar, S. S., Corrin, B., Baak, J. P., Berkel, J., and Vanderschueren, R. G. Second primary lung cancer: importance of long term follow-up. *Thorax* 44: 788–793, 1989.
- Boice, J. D., and Fraumeni, J. F. Second cancer following cancer of the respiratory system in Connecticut, 1935–1982. *J. Natl. Cancer Inst. Monogr.* 68: 83–98, 1985.
- Pairero, P. C., Williams, D. E., Bergsrahl, E. J., Pichter, J. M., Bernatz, P. E., and Payne, N. J. Postsurgical stage I bronchogenic carcinoma. Morbid implications of recurrent disease. *Ann. Thorac. Surg.* 38: 331–338, 1984.
- Shields, T. W., Humphrey, R. W., Higgins, G. A., and Keehn, R. J. Long term survivors after resection of lung carcinoma. *J. Thorac. Cardiovasc. Surg.* 76: 439–442, 1978.
- Auerbach, O., Stout, A. P., Hammond, E. C., and Garfinkel, L. Changes in bronchial epithelium in relation to cigarette smoking and in relation to lung cancer. *N. Engl. J. Med.* 276: 111–118, 1962.
- Auerbach, O., Hammond, E. C., and Garfinkel, L. Changes in bronchial epithelium in relation to cigarette smoking, 1955–1960 vs. 1970–1977. *N. Engl. J. Med.* 300: 381–386, 1979.
- Sundaresan, V., Ganly, P., Haslett, P., Rudd, R., Sinha, G., Bleehen, N. M., and Rabbitts, P. p53 and chromosome 3 abnormalities, characteristic of malignant lung tumors, are detectable in preinvasive lesions of the bronchus. *Oncogene* 7: 1989–1997, 1992.
- Sozzi, G., Miozzo, M., Donghi, R., Pilotti, S., Cariani, C. T., Pastorino, U., Pianta, G. P., and Pierotti, M. A. Deletions of 17p and p53 mutations in preneoplastic lesions of the lung. *Cancer Res.* 52: 6079–6082, 1992.
- Bennett, W. P., Colby, T. V., Travis, W. D., Borkowski, A., Jones, R. T., Lane, D. P., Metcalf, R. A., Samet, J. M., Takeshima, Y., Gu, J. R., Vähäkangas, K. H., Soini, N., Pääkkö, P., Welsh, J. A., Trump, B. F., and Harris, C. C. p53 protein accumulates frequently in early bronchial neoplasia. *Cancer Res.* 53: 4817–4822, 1993.
- Sozzi, G., Miozzo, M., Pastorino, U., Pilotti, S., Donghi, R., Giarola, M., Gregorio, L. D., Manenti, G., Radice, P., Minoretto, F., Porta, G. D., and Pierotti, M. A. Genetic evidence for an independent origin of multiple preneoplastic and neoplastic lung lesions. *Cancer Res.* 55: 135–149, 1995.
- Sozzi, G., Miozzo, M., Tagliabue, E., Calderone, C., Lombardi, L., Pilotti, S., Pastorino, U., Pierotti, M. A., and Porta, G. D. Cytogenetic abnormalities and overexpression of receptors for growth factors in normal bronchial epithelium and tumor samples of lung cancer patients. *Cancer Res.* 51: 400–404, 1991.
- Campbell, A. M., Chavez, P., Vignola, A. M., Bousquet, J., Couret, L., Michel, F. B., and Godard, P. H. Functional characteristics of bronchial epithelium obtained by brushing from asthmatic and normal subjects. *Am. Rev. Respir. Dis.* 147: 529–534, 1993.
- Testa, J., and Siegfried, J. M. Chromosome abnormalities in human non-small cell lung cancer. *Cancer Res.* 52 (Suppl.): 2702–2706, 1992.
- Matturri, L., and Lavezzi, A. M. Recurrent chromosome alterations in non-small cell lung cancer. *Eur. J. Histochem.* 38: 53–58, 1994.
- Reichmann, A., Martin, P., and Levin, B. Karyotypic findings in a colonic villous adenoma. *Cancer Genet. Cytogenet.* 7: 51–57, 1982.
- Moertel, C. A., DeWald, G. W., Coffey, R. J., and Gordon, H. Cytogenetic examination of colonic mucosa in familial polyposis. In: *Proceedings of the Second International Conference on Chromosomes in Solid Tumors*. Tucson, Arizona Cancer Center, pp. 41–48. University of Arizona, 1987.
- Lee, J. S., Pathak, S., Hopwood, V., Tomasovic, H., Mullins, T. D., Baker, F. L., Spitzer, G., and Neidhart, J. A. Involvement of chromosome 7 in primary lung tumor and nonmalignant normal lung tissue. *Cancer Res.* 47: 6349–6352, 1987.
- Ishikawa, Y., Nakagawa, K., Satoh, Y., Kitagawa, T., Sugano, H., Hirano, T., and Tsuchiya, E. Hot spots of chromosomal accumulation at bifurcations of chromate workers' bronchi. *Cancer Res.* 54: 2342–2346, 1994.
- Lechner, J. F., and LaVeck, M. A. A serum-free method for culturing normal human bronchial epithelial cells at clonal density. *J. Tissue Culture Methods* 9: 43–48, 1985.
- Saccomanno, G. *Pulmonary Cytology*. Ed. 2. Chicago: American Society of Clinical Pathologists Press, 1986.
- Lee, J. S., Lippman, S. M., Hong, W. K., Ro, J. Y., Kim, S. Y., Lotan, R., and Hittelman, W. N. Determination of biomarkers for intermediate end points in chemoprevention. *Cancer Res.* 52 (Suppl): 2702s–2710s, 1992.
- United States Department of Health and Human Services. Reducing the Health Consequences of Smoking: 25 Years of Progress. A Report of the Surgeon General. Department of Health and Human Services Publication No. (CDC) 89–8411. W. Services, Pub. Disease Prev.
- National Effects of the Deposited or 1988.
- Lubin, J. E., Kushiak, J. Woodward, miners and 817–827, 19
- Voravud W. N. In: multistage 11
- Thiervi G. and Pal premalignant 5133–5139.
- Shiseki, Yokota, J. F non-small ce
- Merlo, J. Sidransky, E. Cancer Res.
- Spurr, J. Bodmer, W. homologues
- Rusch, J. of p53 or 11 neoplasia, a Cancer Res.
- Veale, J. growth facts 1987.
- Tadashi K., Rikimur

89-8411. Washington, DC: United States Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 1989.

28. National Research Council. Report of the Committee on the Biological Effects of Ionizing Radiation: Health Effects of Radon and Other Internally Deposited  $\alpha$  Emitters (BEIR IV). Washington DC: National Academy Press, 1988.

29. Lubin, J. H., Boice, J. D., Jr., Edling, C., Hornung, R. W., Howe, G. R., Kunz, L., Kusiak, R. A., Morrison, H. L., Radford, E. P., Samet, J. M., Tirmarche, M., Woodward, A., Yao, S. X., and Pierce, D. A. Lung cancer in radon-exposed miners and estimation of risk from indoor exposure. *J. Natl. Cancer Inst.* 87: 817-827, 1995.

30. Voravud, N., Shin, D. M., Ro, J. Y., Lee, J. S., Hong, W. K., and Hittelman, W. N. Increased polysomies of chromosomes 7 and 17 during head and neck multistage tumorigenesis. *Cancer Res.* 53: 2874-2883, 1993.

31. Thiberville, L., Payne, P., Vielkind, J., LeRiche, J., Horsman, D., Nouvet, G., and Palcic, B. Evidence of cumulative gene losses with progression of premalignant epithelial lesions to carcinoma of the bronchus. *Cancer Res.* 55: 5133-5139, 1995.

32. Shiseki, M., Kohno, T., Nishikawa, R., Sameshima, Y., Mizoguchi, H., and Yokota, J. Frequent allelic loss on chromosomes 2q, 18q, and 22q in advanced non-small cell lung carcinoma. *Cancer Res.* 54: 5643-5648, 1994.

33. Merlo, A., Mabry, M., Gabrielson, E., Vollmer, R., Baylin, S. B., and Sidransky, D. Frequent microsatellite instability in primary small cell lung cancer. *Cancer Res.* 54: 2098-2101, 1994.

34. Spurr, N. K., Solomon, E., Jansson, M., Sheen, D., Goodfellow, P. N., Bodmer, W. F., and Verastrom, B. Chromosomal localization of the human homologues to the oncogenes *erb-A* and *B*. *EMBO J.* 3: 159-164, 1984.

35. Rusch, V., Klimstra, D., Linkov, L., and Dmitrovsky, E. Aberrant expression of p53 or the epidermal growth factor receptor is frequent in early bronchial neoplasia, and coexpression precedes squamous cell carcinoma development. *Cancer Res.* 55: 1365-1372, 1995.

36. Veale, D., Ashcroft, T., March, C., Gibson, G. J., and Harris, A. L. Epidermal growth factor receptors in non-small cell lung cancer. *Br. J. Cancer* 55: 513-516, 1987.

37. Tadashi, Y., Kamata, N., Kawano, H., Shimizu, S., Kuroki, T., Toyoshima, K., Rikimura, K., Nomura, N., Ishizaki, R., Pastan, I., Gambou, J., and Shimizu, N.

High incidence of amplification of the epidermal growth factor receptor gene in human squamous carcinoma cell lines. *Cancer Res.* 46: 414-416, 1986.

38. Shin, D. M., Ro, J. Y., Hong, W. K., and Hittelman, W. N. Dysregulation of epidermal growth factor receptor expression in premalignant lesions during head and neck tumorigenesis. *Cancer Res.* 54: 3153-3159, 1994.

39. Soschek, C. M., and King, L. B. Functional and structural characteristics of EGF and its receptor and their relationship to transforming proteins. *J. Cell. Biochem.* 31: 135-152, 1986.

40. Prat, M., Narsimhan, R. P., Crepaldi, T., Nicotra, M. R., Natali, P. G., and Comoglio, P. M. The receptor encoded by the human *c-met* oncogene is expressed in hepatocytes, epithelial cells, and solid tumors. *Int. J. Cancer* 49: 323-328, 1991.

41. Liu, C., and Tsao, M-S. *In vitro* and *in vivo* expression of transforming growth factor  $\alpha$  and tyrosine kinase receptors in human non-small cell lung carcinoma cell lines. *Am. J. Pathol.* 142: 1155-1162, 1993.

42. Giordano, S., Ponzetto, C., Di Renzo, M. F., Cooper, S., and Comoglio, P. M. Tyrosine kinase receptor indistinguishable from the *c-met* protein. *Nature (Lond.)* 339: 155-156, 1989.

43. Naldini, L., Vigna, E., Narsimhan, R. P., Gandino, G., Zarnegar, R., Michalopoulos, G. K., and Comoglio, P. M. Hepatocyte growth factor (HGF) stimulates the tyrosine kinase activity of the receptor encoded by the proto-oncogene *c-MET*. *Oncogene* 6: 501-504, 1991.

44. Liu, C., and Tsao, M-S. Proto-oncogene and growth factor/receptor expression in the establishment of primary human non-small cell lung carcinoma cell lines. *Am. J. Pathol.* 142: 413-423, 1991.

45. Kishimoto, Y., Sugio, K., Hung, J. Y., Virmani, A. K., McIntire, D. D., Minna, J. D., and Gazdar, A. F. Allele-specific loss in chromosome 9p loci in preneoplastic lesions accompanying non-small-cell lung cancers. *J. Natl. Cancer Inst.* 87: 1224-1229, 1995.

46. Hung, J., Kishimoto, Y., Sugio, K., Virmani, A., McIntire, D. D., Minna, J. D., and Gazdar, A. F. Allele-specific chromosome 3p deletions occur at an early stage in the pathogenesis of lung carcinoma. *JAMA* 273: 558-563, 1995.

47. Pastorino, U., Sozzi, G., Miozzo, M., Tagliabue, E., Pilotti, S., and Pierotti, M. A. Genetic changes in lung cancer. *J. Cell. Biochem.* 17F (Suppl.): 237-248, 1993.

# From genes to protein structure and function: novel applications of computational approaches in the genomic era

Jeffrey Skolnick and Jacquelyn S. Fetrow

The genome-sequencing projects are providing a detailed 'parts list' of life. A key to comprehending this list is understanding the function of each gene and each protein at various levels. Sequence-based methods for function prediction are inadequate because of the multifunctional nature of proteins. However, just knowing the structure of the protein is also insufficient for prediction of multiple functional sites. Structural descriptors for protein functional sites are crucial for unlocking the secrets in both the sequence and structural-genomics projects.

**G**enome-sequencing projects are providing a detailed 'parts list' for life. Unfortunately, this list, a portion of which represents the amino acid sequence of all the proteins in a given genome, does not come with an instruction manual. That is, given the genome's sequences, one does not necessarily know straight away which regions encode proteins, which serve a regulatory role and which are responsible for the structure and replication of the DNA itself.

This is not unlike giving a child a list of parts necessary to create a working automobile. Without the necessary expertise, creating the final, working car from just the initial parts list is a nearly impossible task. Similarly, understanding how to create a complete, functioning cell given just the sequence of nucleotides found in an organism's genome is a complex problem.

## What is a protein function?

After a genome is sequenced and its complete parts list determined, the next goal is to understand the function(s) of each part, including that of the proteins. What do we mean by protein function, the focus of this article?

Function has many meanings. At one level, the protein could be a globular protein, such as an enzyme, hormone or antibody, or it could be a structural or membrane-bound protein. Another level is its biochemical function, such as the chemical reaction and the substrate specificity of an enzyme. The regulatory molecules or cofactors that bind to a protein are also levels of biochemical function.

At the cellular level, the protein's function would involve its interaction with other macromolecules and the function and cellular location of such complexes. There is also the protein's physiological function; that is, in which metabolic pathway the protein is involved or what physiological role it performs in the organism. Finally, the phenotypic function is the role played by the protein in the total organism, which is observed by deleting or mutating the gene encoding the protein.

Obviously, the complete characterization of protein function is difficult but efforts are under way at all levels<sup>1-4</sup>, including cellular function<sup>5,6</sup>. In this article, however, we focus on identifying the biochemical function of a protein given its sequence, a problem that is amenable to molecular approaches.

## Sequence-based approaches to function prediction

The sequence-to-function approach is the most commonly used function-prediction method. This robust field is well developed and, in the interest of space limitations, we will merely present a brief overview.

There are two main flavors of this approach: sequence alignment<sup>7-9</sup>; and sequence-motif methods such as Prosite<sup>10</sup>, Blocks<sup>11</sup>, Prints<sup>12,13</sup> and Emotif<sup>14</sup>. Both the alignment and the motif methods are powerful but a recent analysis has demonstrated their significant limitations<sup>15</sup>, suggesting that these methods will increasingly fail as the protein-sequence databases become more diverse.

An extension of these approaches that combines protein-sequence with structural information has been developed and some successes have been reported<sup>16</sup>. However, this method still applies the structural information in a one-dimensional, 'sequence-like' fashion and fails to take into account the powerful three-dimensional information displayed by protein structures.

In addition, proteins can gain and lose function during evolution and may, indeed, have multiple functions in the cell (Box 1). Sequence-to-function methods cannot specifically identify these complexities. Inaccurate use of sequence-to-function methods has led to significant function-annotation errors in the sequence databases<sup>17</sup>.

## An alternative approach

An alternative, complementary approach to protein function prediction uses the sequence-to-structure-to-function paradigm. Here, the goal is to determine the structure of the protein of interest and then to identify the functionally important residues in that structure. Using the chemical structure itself to identify functional sites is more in line with how the protein actually works.

J. Skolnick (skolnick@danforthcenter.org) is at the Danforth Plant Science Center, Laboratory of Computational Genomics, 4041 Forest Park Avenue, St. Louis, MO 63108, USA. J.S. Fetrow is at GeneFormatics, Suite 200, 5830 Oberlin Drive, San Diego, CA 92121-3754, USA.

In a sense, this is one long-term goal of 'structural genomics' projects<sup>18,19</sup>, which are designed to determine all possible protein folds experimentally, just as genome-sequencing projects are determining all

structural-biology approaches, in which one knows the protein's function first and only then, if the function is sufficiently important, determines its structure.

It is implicitly assumed that having the protein's structure will provide insights into its function, thereby furthering the goals of the human-genome-sequencing project. However, knowing a protein's three-dimensional structure is insufficient to determine its function (Box 2). What we really need to analyse and predict the multifunctional aspects of proteins is a method specifically to recognize active sites and binding regions in these protein structures.

#### Active-site identification

In order to use a structure-based approach to function prediction, one must identify the key residues responsible for a given biochemical activity. For many years, it has been suggested that the active sites in proteins are better conserved than the overall fold. Taken to the limit, this suggests that one could not only identify distant ancestors with the same global fold and the same activity but also proteins with similar functions but distantly related, or possibly unrelated, global folds.

The validity of this suggestion was demonstrated empirically by Nussinov and co-workers, who showed that the active sites of eukaryotic serine proteases, subtilisins and sulphhydryl proteases exhibit similar structural motifs<sup>21</sup>. Furthermore, in a recent modeling study of *Saccharomyces cerevisiae* proteins, protein functional sites were found to be more conserved than other parts of the protein models<sup>22</sup>. Similarly, it has been demonstrated that the catalytic triad of the  $\alpha/\beta$  hydrolases is structurally better conserved than other histidine-containing triads<sup>23</sup>. A comparison of the structure of the hydrolase catalytic triad to other histidine-containing triads shows a distinct bimodal distribution, while a similar analysis done with a randomly selected triad shows a unimodal distribution (Fig. 1).

Kasuya and Thornton<sup>24</sup> generalized this example by creating structural analogs of a few Prosite sequence motifs<sup>10</sup>. For the 20 most-frequently occurring Prosite patterns, the associated local structure is quite distinct. These results provide clear evidence that enzyme active sites are indeed more highly conserved than other parts of the protein.

#### Identifying active sites in experimental structures

Historically, several groups have attempted to identify functional sites in proteins; these efforts were directed at protein engineering or building functional sites in places where they did not previously exist. This has been successfully accomplished for several metal-binding sites<sup>25-33</sup>. However, highly accurate functional-site descriptors of the backbone and side-chain atoms were required, fueling the belief that significant atomic detail is required in site descriptors for function identification.

Highly detailed residue side-chain descriptors of the active sites of serine proteases and related proteins have been used to identify functional sites<sup>3</sup>. The use of these highly detailed motifs has led to the identification of

### Box 1. Proteins are multifunctional

A common protein characteristic that makes functional analysis based on local structure especially difficult is the tendency of proteins to be multifunctional. For instance, a catalytic site requires histidine, serine and zinc, and performs a redox reaction. Each of these occurs at different functional sites that are in close proximity and the combination of all four sites creates the fully functional protein.

Other examples of multifunctional proteins are the nucleic-acid-binding proteins. For instance, DNA regulatory proteins often contain a DNA-binding domain, a multimerization domain and additional sites that bind regulatory proteins; a classic example is RecA<sup>59</sup>. The 3C rhinovirus protease exhibits a proteolytic function as well as an RNA-binding function<sup>60,61</sup>. Transcription factors are also complex, multifunctional proteins<sup>62</sup>. It is becoming increasingly important to recognize each of these different functions of gene products of a newly sequenced gene.

The serine-threonine-phosphatase superfamily is a prime example of the difficulties of using standard sequence analysis to recognize the multiple functions found in single proteins. This large protein family is divided into a number of subfamilies, all of which contain an essential phosphatase active site. Subfamilies 1, 2A and 2B exhibit 40% or more sequence identity between them<sup>63</sup>. However, each of these subfamilies is apparently regulated differently in the cell<sup>64-67</sup> and observation suggests that there are different functional sites at which regulation can occur. Because the sequence identity between subfamilies is so high, standard sequence-similarity methods could easily misclassify new sequences as members of the wrong subfamily if the functional sites are not carefully considered, as was recently demonstrated<sup>43</sup>.

These are but a few examples of the multifunctionality of proteins. The recognition of this multifunctional nature is of critical importance to the genomics field. Useful functional-annotation methods must consider all of the specific functions in a given protein and will not just provide a general classification of function.

several novel functional sites in known, high-quality protein structures<sup>3,34</sup>. More automated methods for finding spatial motifs in protein structures have also been described<sup>21,34-40</sup>.

Unfortunately, most of these methods require the exact placement of atoms within protein backbones and side chains, and so have not been shown to be relevant to inexact predicted structures. Recently, however, we described the production of fuzzy, inexact descriptors of protein functional sites<sup>15</sup>. As we wish to apply the descriptors to experimental structures as well as to predicted protein models, we used only carbon atoms and side-chain centers-of-mass positions. We call these descriptors 'fuzzy functional-forms' (FFFs) and have created them for both the disulfide-oxidoreductase<sup>15,41</sup> and  $\alpha/\beta$ -hydrolase catalytic active sites<sup>23</sup>.

The disulfide-oxidoreductase FFF was applied to screen high-resolution structures from the Brookhaven protein database<sup>42</sup>. In a dataset of 364 protein structures, the FFF accurately identified all proteins known to exhibit the disulfide-oxidoreductase active site<sup>15</sup>. In a larger dataset of 1501 proteins, the FFF again accurately identified all proteins with the active site. In addition, it identified another protein, 1fjm, a serine-threonine phosphatase. This result was initially discouraging but subsequent sequence alignment and clustering analysis strongly suggested that this putative site might indeed be a site of redox regulation in the serine-threonine phosphatase-1 subfamily<sup>43</sup>. If confirmed by experiment, this result will highlight the advantages of using structural descriptors to analyse multiple functional sites in proteins. It will also highlight the fact that human

## Box 2. Knowing a protein's structure does not necessarily tell you its function

Because proteins can have similar folds but different functions<sup>69</sup>, determining the structure of a protein may or may not tell you something about its function. The most well-studied example is the  $(\alpha/\beta)_8$  barrel enzymes, of which triose-phosphate isomerase (TIM) is the archetypal representative. Members of this family have similar overall structures but different functions, including different active sites, substrate specificities and cofactor requirements<sup>70,71</sup>.

Is this example common? Our own analysis of the 1997 SCOP database<sup>68</sup> shows that the five largest fold families are the ferredoxin-like, the  $(\alpha/\beta)$  barrels, the knottins, the immunoglobulin-like and the flavodoxin-like fold families with 22, 18, 13, 9 and 9 superfamilies, respectively (Fig. 1). In fact, 57 of the SCOP fold families consist of multiple superfamilies. These data only show the tip of the iceberg, because each superfamily is further composed of protein families and each individual family can have radically different functions. For example, the ferredoxin-like superfamily contains families identified as Fe-S ferredoxins, ribosomal proteins, DNA-binding proteins and phosphatases, among others.

After this article was submitted, a much more detailed analysis of the SCOP database was published<sup>72</sup>. This finds a broad function-structure correlation for some structural classes, but also finds a number of ubiquitous functions and structures that occur across a number of families. The article provides a useful analysis of the confidence with which structure and function can be correlated<sup>72</sup>. Knowing the protein structure by itself is insufficient to annotate a number of functional classes and is also insufficient for annotating the specific details of protein function.

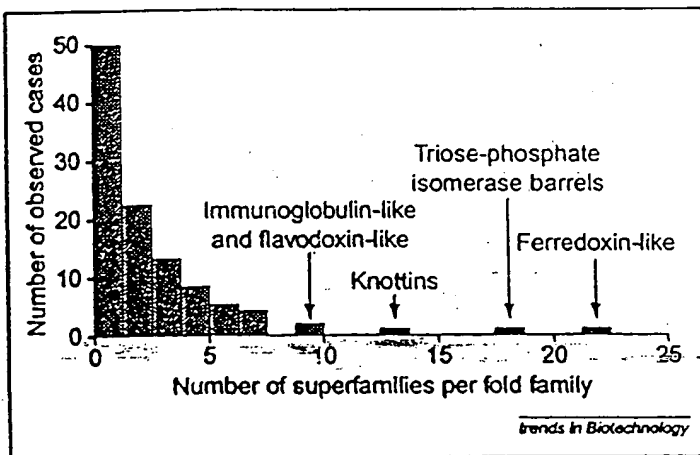


Figure 1

Histogram of the numbers of superfamilies found in each SCOP fold family. These data clearly show that proteins with similar structures can have different functions and demonstrate the difficulty of assigning protein function based simply on the three-dimensional structure. The data were taken from the 1997 distribution of SCOP (<http://scop.mrc-lmb.cam.ac.uk/scop>). For a more detailed analysis, see Ref. 72.

observation alone is no longer adequate for identifying all functional sites in known protein structures.

To date, the use of structure to identify function has largely focused on high-resolution structures and highly detailed descriptors of protein functional sites. However, the creation of inexact descriptors for functional sites opens the way to the application of these methods to inexact, predicted protein models. The question remains: how good does a model have to be in order to use FFFs to identify its active sites?

## The state of the art in structure-prediction methods

For proteins whose sequence identity is above ~30%, one can use homology modeling to build the structure of a protein from a known structure. This is difficult for proteins that are not homologous to proteins with known structure. At present, there are two approaches for these sequences: *ab initio* folding<sup>45-48</sup> and threading<sup>49-53</sup>.

In *ab initio* folding, one starts from a random conformation and then attempts to assemble the native structure. As this method does not rely on a library of pre-existing folds, it can be used to predict novel folds. The recent CASP3 protein-structure-prediction experiment (<http://PredictionCenter.Unl.gov/CASP3>) involved the blind prediction of the structure of proteins whose actual structure was about to be experimentally determined. These results indicate that considerable progress has been made<sup>46,54</sup>. For helical and  $\alpha/\beta$  proteins with less than 110 residues, structures were often predicted whose backbone root-mean-square deviation (RMSD) from native ranged from 4-7 Å. Progress is being made with the  $\beta$  proteins, too, although they remain problematic. Because *ab initio* methods can identify novel folds, these methods could be used to help to select sequences likely to yield novel folds in experimental structural-genomics projects.

Another approach to tertiary-structure prediction is threading. Here, for the sequence of interest, one attempts to find the closest matching structure in a library of known folds<sup>52,55</sup>. Threading is applicable to proteins of up to 500 residues or so and is much faster than *ab initio* approaches. However, threading cannot be used to obtain novel folds.

## Ab initio predicted models can be used for automatic protein-function prediction

The results of the recent CASP3 competition suggest that current modeling methods can often (but not always) create inexact protein models. Are these structures useful for identifying functional sites in proteins? Using the *ab initio* structure-prediction program MONSTER, the tertiary structure of a glutaredoxin, 1ego, was predicted<sup>56</sup>. For the lowest-energy model, the overall backbone RMSD from the crystal structure was 5.7 Å.

To determine whether this inexact model could be used for function identification, the sets of correctly and incorrectly folded structures were screened with the FFF for disulfide-oxidoreductase activity<sup>15</sup>. The FFF uniquely identified the active site in the correctly folded structure but not in the incorrectly folded ones (Fig. 2). This is a proof-of-principle demonstration that inexact models produced by *ab initio* prediction of structure from sequence can be used for the subsequent prediction of biochemical function. Of course, improvements in the method have to be made before such predictions can be done on a routine basis.

## Use of predicted structures from threading in protein-function prediction

At present, practical limitations preclude folding a entire genome of proteins using *ab initio* methods<sup>5</sup>. Threading is more appropriate for achieving the requisite high-throughput structure prediction. Thus, a standard threading algorithm<sup>58</sup> has been used to screen

proteins in nine genomes for the disulfide-oxidoreductase active site described above.

First, sequences that aligned with the structures of known disulfide oxidoreductases were identified. Then, the sequences were aligned to the disulfide-oxidoreductase active site residues and geometry. For those sequences for which other homologs were available, a sequence-conservation profile was constructed<sup>23</sup>. If the putative active-site residues were not conserved in the sequence subfamily to which the protein belongs, that sequence was eliminated. Otherwise, the sequence is predicted to have the function.

Using this sequence-to-structure-to-function method, 99% of the proteins in the nine genomes that have known disulfide-oxidoreductase activity have been found. From 10% to 30% more functional predictions are made than by alternative sequence-based approaches; similar results are seen for the  $\alpha/\beta$  hydrolases<sup>23</sup>. Surprisingly, in spite of the fact that threading algorithms have problems generating good sequence-to-structure alignments, active sites are often accurately aligned, even for very distant matches. This observation would agree with the above experimental results indicating that active sites are well conserved in protein structures.

Importantly, the false-positive rate when using structural information is much lower than that found using sequence-based approaches, as demonstrated by a detailed comparison of the FFF structural approach and the Blocks sequence-motif approach (N. Siew *et al.*, unpublished). In this study, the sequences in eight genomes, including *Bacillus subtilis*, were analysed for disulfide-oxidoreductase function using the disulfide-oxidoreductase FFF, the thioredoxin Block 00194 and the glutaredoxin Block 00195. If we assume that those sequences identified by both the FFF and Blocks are 'true positives', we find 13 such sequences in the *B. subtilis* genome.

There is no experimental evidence validating all of these 'true positives' and so they are more accurately termed 'consensus positives'. In order to find these 13 'consensus positive' sequences, the FFF hits seven false positives. On the other hand, Blocks hits 23 false positives (Fig. 3). It was previously suggested that the use of a functional requirement adds information to threading and reduces the number of false positives<sup>52</sup>. These data, including the data shown in Fig. 3, validate this claim on a genome-wide basis.

Of course, as no genome has had the function of all of its proteins experimentally annotated, it is impossible to know how many other proteins with the specified biochemical function were not properly identified. This is a critical question for researchers attempting to predict protein function. Experimental confirmation will be needed to validate this or any other method fully. This points out the need for closely coupling computational function-prediction algorithms with experiments.

#### Weaknesses of using the sequence-to-structure-to-function method of function prediction

Based on studies to date, the identification of enzymatic activity requires a model in which the backbone RMSD from native near the active sites is about 4–5 Å. Predicted models are better at describing the geometry in the core of the molecule than in the loops and so

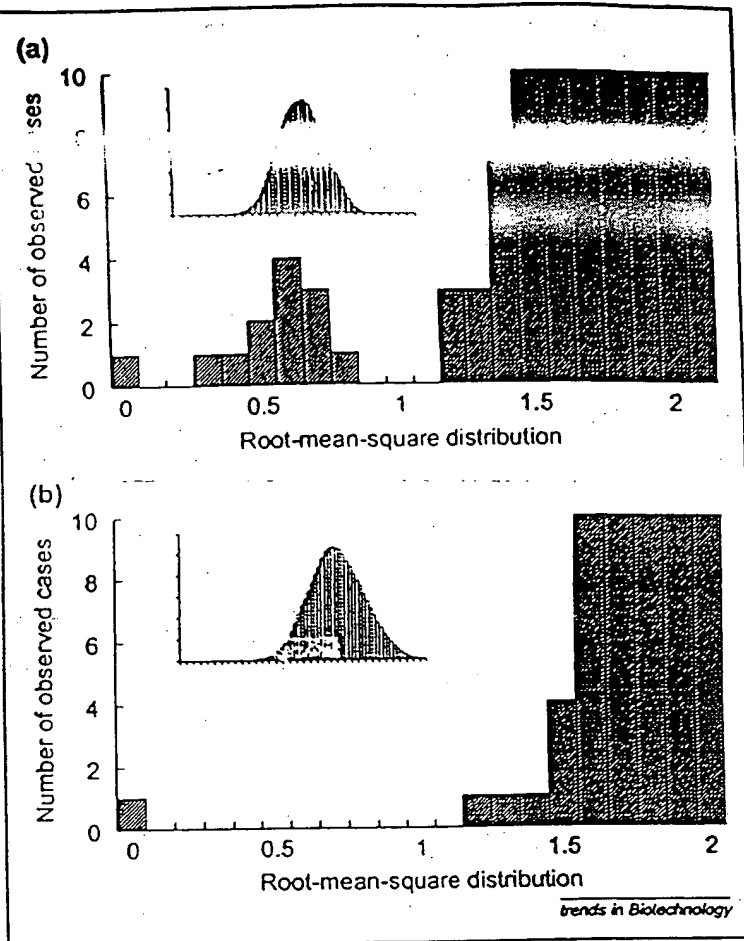


Figure 1

The distribution of root-mean-square distributions (RMSD) between the hydrolase catalytic triad and all other histidine-containing triads shows a bimodal distribution (a); by contrast, the RMSD between a randomly selected (non-catalytic) triad and all other histidine-containing triads has a unimodal distribution (b). The His-Ser-Asp catalytic triad in the protein-1 gpl (Rp2 lipase) (a) and a random histidine-containing triad from 4pga (glutaminase-asparaginase) (b) were structurally aligned to all His-containing triads in a database of 1037 proteins<sup>23</sup>. Actual  $\alpha/\beta$ -hydrolase active sites (a) and the 4pga site (b) are indicated by blue bars; other histidine triads that are not active sites are indicated by red bars. None of the sites found by matching to the 4pga were hydrolase active sites. Inset graphs show the full distribution.

predicting the function of a protein whose active site is in loops may be a problem. Also, the method can currently only be applied to enzyme active sites; substrate- and ligand-binding sites have not been identified using the inexact models. Techniques that will further refine inexact protein models will be quite useful in taking the protein analysis to the next step.

#### Conclusions

Although sequence-based approaches to protein-function prediction have proved to be very useful, alternatives are needed to assign the biochemical function of the 30–50% of proteins whose function cannot be assigned by any current methods. One emerging approach involves the sequence-to-structure-to-function paradigm. Such structures might be provided by structural-genomics projects or by structure-prediction algorithms. Functional assignment is made by screening the resulting structure against a library of structural descriptors for known active sites or binding regions.

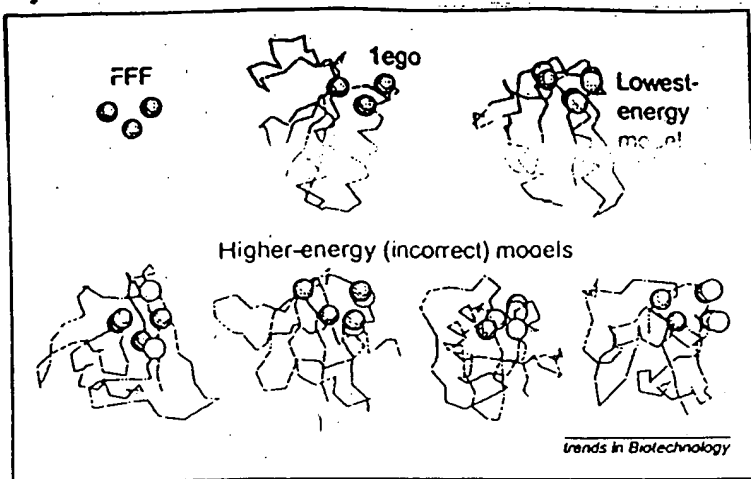


Figure 2

Application of the disulfide oxidoreductase fuzzy functional form (FFF) to *ab initio* models of glutaredoxin created by the program MONSTER shows that the FFF can distinguish between correctly folded and misfolded (or higher-energy) models. The FFF is shown as two orange balls (representing the cysteines) and a blue ball (representing the proline). The protein models are shown as magenta wire models with the active-site cysteines and proline shown as yellow and cyan balls, respectively. The FFF clearly distinguishes the correct active site in the crystal structure of the glutaredoxin 1ego and the correctly folded, lowest-energy model. The FFF does not match to the active sites of any of the higher energy, misfolded structures, four of which are shown here.

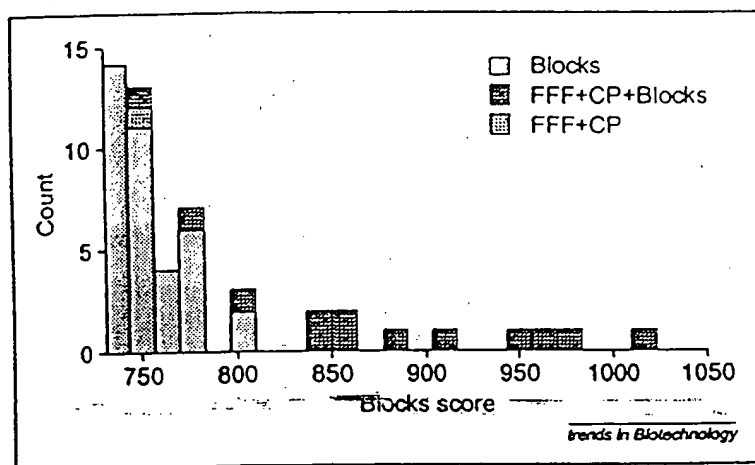


Figure 3

Analysis of the *Bacillus subtilis* genome using the thioredoxin Block 00194. The Blocks score (computed using the publicly available BLUMPS program) is plotted on the x axis and the number of sequences found in each scoring bin is plotted on the y axis. Those sequences identified as 'consensus positives' (identified by both the fuzzy functional form (FFF) and the Block) are shown as red bars. One additional sequence found by the FFF, which is likely to be a true positive, is shown as a blue bar. All other sequences, putative 'false positives', are shown as yellow bars. Using the Blocks score at which all 13 of the 'consensus positives' are found, 23 false positives are also found. In its analysis of the *B. subtilis* genome, the FFF identifies only seven false positives along with the same 13 'consensus positives' (data not shown).

Detailed descriptors will only work on the experimentally determined, high-quality structures. Ideally, however, the descriptors should work on both experimental structures and the cruder models provided by tertiary-structure-prediction algorithms.

The advantages of such an approach are that one need not establish an evolutionary relationship in order to assign function, that more than one function can be

assigned to a given protein [an issue of major importance, because proteins are multifunctional (Box 1)] and, ultimately, that having a structure can provide deeper insight into the biological mechanism of protein function and regulation. The disadvantages are that one needs to have the protein's structure before a function can be assigned and that the approach is limited to those functions associated with proteins with at least one solved structure, so that a functional-site descriptor can be constructed.

In this sense, structure-to-function assignment can be thought of as 'functional threading' – find the active-site match in a library of descriptors for known protein active sites. This is the first step in the long process of using structure to assign all levels of function, a goal that is made increasingly important with the emergence of structural genomics. Based on the progress to date, it is apparent that structure will play an important role in the post-genomic era of biology.

### Acknowledgment

We thank L. Zhang for producing the data in Box 2 and Fig. 1.

### References

- Gird, F.R.N. and Rohlfsch, T.M. (1979) Motions in proteins. *Adv Protein Chem.* 33, 73–165
- Laskowski, R.A. et al. (1996) X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. *J. Mol. Biol.* 259, 175–201
- Wallace, A.C. et al. (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* 5, 1001–1013
- Henikoff, S. and Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* 19, 6565–6572
- Riley, M. (1993) Functions of gene products of *Escherichia coli*. *Microbiol. Rev.* 57, 862–952
- Karp, P.D. and Riley, M. (1993) Representations of metabolic knowledge. *Ismb* 1, 207–215
- Altschul, S.F. et al. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
- Pearson, W.R. (1996) Effective protein sequence comparison. *Methods Enzymol.* 266, 227–258
- Sturrock, S.S. and Collins, J.F. (1993) *Biocomputing Research Unit*. University of Edinburgh, Edinburgh, UK
- Bairoch, A. et al. (1995) The PROSITE database, its status in 1995. *Nucleic Acids Res.* 24, 189–196
- Henikoff, S. and Henikoff, J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics* 19, 97–107
- Attwood, T.K. et al. (1994) PRINTS – A database of protein motif fingerprints. *Nucleic Acids Res.* 22, 3590–3596
- Attwood, T.K. et al. (1997) Novel developments with the PRINTS protein fingerprint database. *Nucleic Acids Res.* 25, 212–216
- Nevill-Manning, C.G. et al. (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5865–5871
- Petrov, J.S. and Skolnick, J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* 281, 949–968
- Yu, L. et al. (1998) A homology identification method that combine protein sequence and structure information. *Protein Sci.* 7, 2499–2510
- Bork, P. and Bairoch, A. (1996) Go hunting in sequence database but watch out for traps. *Trends Genet.* 12, 425–427
- Gasterland, T. (1998) Structural genomics: bioinformatics in the driver's seat. *Nat. Biotechnol.* 16, 625–627
- McKusick, V.A. (1997) Genomics: structural and functional studies of genomes. *Genomics* 45, 244–249
- Montelione, G.T. and Anderson, S. (1999) Structural genomics: a keystone for a human proteomic project. *Nat. Struct. Biol.* 6, 11–1



- 21 Fischer, D. *et al.* (1994) Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci.* 3, 769-778
- 22 Sanchez, P. and Sali, A. (1998) Large-scale protein structure modeling of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 13597-13602
- 23 Zhang, L. *et al.* (1998) Functional analysis of *E. coli* proteins for members of the  $\alpha/\beta$  hydrolase family. *Fold. Design* 3, 535-548
- 24 Kasuya, A. and Thornton, J.M. (1999) Three-dimensional structure analysis of Prosite patterns. *J. Mol. Biol.* 286, 1673-1691
- 25 Coldren, C.D. *et al.* (1997) The rational design and construction of a cuboidal iron-sulfur protein. *Proc. Natl. Acad. Sci. U. S. A.* 94, 6635-6640
- 26 Pinto, A.L. *et al.* (1997) Construction of a catalytically active iron superoxide dismutase by rational protein design. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5562-5567
- 27 Hellings, H.W. and Richards, F.M. (1991) Construction of new ligand binding sites in proteins of known structure: (I) computer-aided modeling of sites with pre-defined geometry. *J. Mol. Biol.* 222, 763-785
- 28 Hellings, H.W. *et al.* (1991) Construction of new ligand binding sites in proteins of known structure: (II) grafting of a buried transition metal binding site into *Escherichia coli* thioredoxin. *J. Mol. Biol.* 222, 787-803
- 29 Klembe, M. and Regan, L. (1995) Characterization of metal binding by a designed protein: single ligand substitutions at a tetrahedral Cys<sub>4</sub>His<sub>2</sub> site. *Biochemistry* 34, 10094-10100
- 30 Klembe, M. *et al.* (1995) Novel metal-binding proteins by design. *Nat. Struct. Biol.* 2, 368-373
- 31 Farinas, E. and Regan, L. (1998) The *de novo* design of a rubredoxin-like Fe site. *Protein Sci.* 7, 1939-1946
- 32 Crowder, M.W. *et al.* (1995) Spectroscopic studies on the designed metal-binding sites of the 43C9 single chain antibody. *J. Am. Chem. Soc.* 117, 5627-5634
- 33 Halfon, S. and Craik, C.S. (1996) Regulation of proteolytic activity by engineered tridentate metal binding loops. *J. Am. Chem. Soc.* 118, 1227-1228
- 34 Wallace, A.C. *et al.* (1997) TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: application to enzyme active sites. *Protein Sci.* 6, 2308-2323
- 35 Kleywegt, G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.* 285, 1887-1897
- 36 Matsuo, Y. and Nishikawa, K. (1994) Protein structural similarities predicted by a sequence-structure compatibility method. *Protein Sci.* 3, 2055-2063
- 37 Russell, R.B. (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* 279, 1211-1227
- 38 Han, K.F. *et al.* (1997) Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci.* 6, 1587-1590
- 39 Artymiuk, P.J. *et al.* (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* 236, 327-344
- 40 Karlin, S. and Zhu, Z.Y. (1996) Characterizations of diverse residue clusters in protein three-dimensional structures. *Proc. Natl. Acad. Sci. U. S. A.* 93, 8344-8349
- 41 Fetrow, J.S. *et al.* (1998) Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/diaredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* 282, 703-711
- 42 Abola, E.E. *et al.* (1987) *Protein Data Bank in Crystallographic Databases: Information Content, Software Systems, Scientific Application* (Allen, F.H. *et al.*, eds), Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester
- 43 Fetrow, J.S. *et al.* (1999) Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J.* 13, 1866-1874
- 44 Sali, A. *et al.* (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins* 23, 314-326
- 45 Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281, 565-577
- 46 Shorrock, D. (1999) The state of the art. *Curr. Biol.* 9, R205-R209
- 47 Lee, J. *et al.* (1999) Calculation of protein conformation by global optimization of a potential energy function. *Protein Sci.* (Suppl.), 204-208
- 48 Ortiz, J. *et al.* (1999) Calculation of protein conformation by restraints derived from evolutionary information. *Protein Sci.* (Suppl.), 177-185
- 49 Bowie, J.U. *et al.* (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164-170
- 50 Finkelstein, A.V. and Reva, B.A. (1991) A search for the most stable folds of protein chains. *Nature* 351, 497-499
- 51 Bryant, S.H. and Lawrence, C.E. (1993) An empirical energy function for threading protein sequence through folding motif. *Proteins* 16, 92-112
- 52 Lathrop, R. and Smith, T.F. (1996) Global optimum protein threading with gapped alignment and empirical pair scoring function. *J. Mol. Biol.* 255, 641-665
- 53 Jones, D.T. *et al.* (1992) A new approach to protein fold recognition. *Nature* 358, 86-89
- 54 Sternberg, M.J. *et al.* (1999) Progress in protein structure prediction assessment of CASP3. *Curr. Opin. Struct. Biol.* 9, 368-373
- 55 Miller, R.T. *et al.* (1996) Protein fold recognition by sequence threading tools and assessment techniques. *FASEB J.* 10, 171-178
- 56 Ortiz, A.R. *et al.* (1998) Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.* 277, 419-448
- 57 Skolnick, J. *et al.* (1998) Reduced protein models and their application to the protein folding problem. *J. Biomol. Struct. Dyn.* 16, 381-396
- 58 Jaroszewski, L. *et al.* (1998) Fold prediction by a hierarchy of sequence, threading and modeling methods. *Protein Sci.* 7, 1431-1440
- 59 Takahashi, M. *et al.* (1996) Locations of functional domains in the RocA protein: overlap of domains and regulation of activities. *Eur. J. Biochem.* 242, 261-268
- 60 Leong, L.E. *et al.* (1993) Human rhinovirus-14 protease 3C (3Cpro) binds specifically to the 5' noncoding region of the viral RNA: evidence that 3Cpro has different domains for the RNA binding and proteolytic activities. *J. Biol. Chem.* 268, 25735-25739
- 61 Matthews, D.A. *et al.* (1994) Structure of human rhinovirus 3C protease reveals a trypsin-like polypeptide fold, RNA-binding site and means for cleaving precursor polypeptide. *Cell* 77, 761-771
- 62 Ladomery, M. (1997) Multifunctional proteins suggest connections between transcriptional and post-transcriptional processes. *BioEssays* 19, 903-909
- 63 Goldberg, J. *et al.* (1995) Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1. *Nature* 376, 745-753
- 64 Murnby, M.C. and Walter, G. (1993) Protein serine/threonine phosphatases: structure, regulation and functions in cell growth. *Physiol. Rev.* 73, 673-699
- 65 Jia, Z. (1997) Protein phosphatases: structures and implications. *Biochem. Cell Biol.* 75, 17-26
- 66 Holmes, C.F.B. and Boland, M.P. (1993) Inhibitors of protein phosphatase-1 and -2A: two of the major serine/threonine protein phosphatases involved in cellular regulation. *Curr. Opin. Struct. Biol.* 3, 934-943
- 67 Neimani, R. and Lee, E.Y.C. (1993) Reactivity of sulphydryl groups of the catalytic subunits of rabbit skeletal muscle protein phosphatases 1 and 2A. *Arch. Biochem. Biophys.* 300, 24-29
- 68 Murzin, A.G. *et al.* (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540
- 69 Orengo, C.A. *et al.* (1997) CATH: a hierarchical classification of protein domain structures. *Structure* 5, 1093-1108
- 70 Lesk, A.M. *et al.* (1989) Structural principles of  $\alpha/\beta$  proteins: the packing of the interior of the sheet. *Protein Struct. Fund. Charact.* 5, 139-148
- 71 Farber, G.K. and Petsko, G.A. (1990) The evolution of  $\alpha/\beta$  barrel enzymes. *Trends Biochem. Sci.* 15, 228-234
- 72 Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* 288, 147-164



# Powers and Pitfalls in Sequence Analysis: The 70% Hurdle

Peer Bork<sup>1</sup>

European Molecular Biology Laboratory (EMBL) 69012 Heidelberg; Germany and Max-Delbrück-Centrum, D-13122 Berlin-Buch, Germany

**H**igh-throughput technologies impress us almost every week with novel global results and big numbers. They often reveal important general trends that are impossible to realize with classical, low-throughput experimental methods, yet (so far) they provide fewer insights into specific, molecular detail. Because of the amount of data involved, high-throughput technologies imply the use of bioinformatics methods that deal with information transformation, storage, and analysis. By necessity, most of these processes are automated.

Partly because of the nature of current publication schemes, the accuracy and error margins of a given method are often only found in small print. It is obvious that each method has its limits and also that during data processing, some information will be lost or diluted. Because of the current need to integrate and add value to data, results from high-throughput experiments (if made publicly accessible) are often taken further by third-party research that relies on the quality of these data. Thus, I believe that public awareness of error margins for high-throughput experimental and computational methods should be increased; the incredibly valuable data accumulating in various heterogeneous databases permit powerful analyses but should not be overinterpreted. In the following discussion, I will concentrate on limits in computational sequence analysis, which is far from being perfect (Table 1), despite the fact that sequencing itself is highly automated and accurate, and despite the fact that sequence information is described in simple linear terms (using a four-letter alphabet). On

average, a 70% accuracy just to predict functional and structural features has to be considered a success (Table 1).

## Limitations in the Total Knowledge Base of Protein Function

As these analysis methods are knowledge based, one of the reasons for the inaccuracy is that the quality of data in public sequence databases is still insufficient (e.g., Bork and Bairoch 1996; Bhatia et al. 1997; Pennisi 1999). This is particularly true for data on protein function. Protein function is loosely defined; cellular function is more than the very complicated network of individual molecular interactions on which it is based (Bork et al. 1998). Furthermore, the semantics for functional features are not always established. For instance, the notion of a "protein complex" not only depends heavily on detection and purification methods—which, in turn, are constantly evolving—but also on environmental conditions. Protein function is context dependent, and both molecular and cellular aspects have to be considered (for review, see Bork et al. 1998).

To illustrate some of this complexity, a good example is lactate dehydrogenase. This gene product can act both as a dehydrogenase and an eye lens structural protein, depending on its context (for review, see Piatigorsky and Wistow 1991). Even without the complication of a second, unrelated role for the same gene product, do we know enough about the function of lactate dehydrogenase, one of the best-studied proteins? We know its biochemical pathway (at least in human and some model organisms), its different isoenzymes (in organisms) with different context-dependent

properties, its regulation, and the organization of its quaternary structure. However, we are probably still missing much information, even on crucial molecular features: Are we sure about alternative splice variants? Can we exclude age-dependent post-translational modifications in some tissues? Our knowledge is even more limited regarding higher order functions that involve concentration, compartmental organization, dynamics, regulation, and perhaps even the impact of external environment. Often, the available data give at best some reliable qualitative results on functional features but far from a complete understanding of functionality. Yet our ability to annotate genome sequences and translate information therein relies heavily on the summaries of features attached to each sequence in the respective public databases.

## Limitations of Gene Expression Data Extrapolations

As more high-throughput technologies follow, the data will become more complicated than sequences. Novel complementary data types such as gene expression arrays will generate more functional information, but conclusions from these data are often stretched with regard to protein products. The expression of genes and their reciprocal proteins seems to correlate weakly, with a correlation coefficient of 0.48 (Anderson and Seilhammer 1997). Furthermore, recent studies (Hanke et al. 1999; Mironov et al. 1999) show that alternative splicing might affect >30% of the human genes, although measurements at the protein level have yet to confirm this. Finally, the number of known post-

<sup>1</sup>E-MAIL bork@embl-heidelberg.de; FAX 11-49-6221-387517.

**Table 1. Selected Examples of Prediction Accuracy in Different Areas of Sequence Analysis**

Prediction of	Acc × cov <sup>a</sup>	Accuracy (%)	Coverage or coverage in % of reference set	Reference <sup>b</sup>
Human promoters	0.35	50	70% of annotated test set	Prestidge 1995; P. Bucher (pers. comm)
Human regulatory RNA elements	0.34	85	40% of new DNA	Dandekar and Sharma (1998)
Human genes (only presence)	0.49	70	70% of chromosome 22	Dunham et al. (1999) and refs. therein
Human SNPs by EST comparison	0.21	70	30% of all proteins with SNP	Buelow et al. (1999); Sunyaev et al. (2000)
Human alternative splicing	0.45	90	50% of all splice sites	Hanke et al. (1999)
Transmembranes (only presence)	0.85	85	99% of annotated test set	Tusnady and Simon (1998) and refs. therein
Signal peptides (only presence)	.90	90	100% of annotated test set	Nielsen et al. (1999)
GPI anchors (incl cleavage site)	.72	72	100% of annotated test set	Eisenhaber et al. (1999)
Coiled coil (only presence)	.81	90	90% of annotated coiled coil	Lupas (1996)
Secondary structure (Three states)	.77	77	100% of 3D test set	Jones (1999) and refs. therein
Buried or exposed residues	.74	74	100% of 3D test set	Rost (1996)
Residue hydration	.72	72	100% of 3D test set	Ehrlich et al. (1998)
Protein folds (in Mycoplasma)	.49	98	50% of Mycoplasma ORFs	Teichmann et al. (1999) and refs. therein
Homology (several methods)	.49	98	50% of 3D test set	Muller et al. (1999) and refs. therein
Functional features by homology	.63	90	70% unicellular genomes	Bork and Koonin (1998); Brenner (1999)
Function association by context	.25	50	10% high confidence in yeast	Marcotte et al. (1999b)
Cellular localization (two states)	.77	77	100% of annotated test set	Andrade et al. (1998)

The numbers referred to are in many cases crude estimates taken or sometimes even estimated from the literature and have an expected accuracy of ~70%. Direct comparison of the numbers might be misleading as the context is not properly explained here. Furthermore, although most of the examples are two state predictions, the percentage numbers do not take into account random occurrences of the states. All test sets are most likely biased (e.g., current 31) test sets do not contain many compositionally biased regions, which probably contain up to 15% of all residues, and annotation test sets are far from being perfect; see text), i.e., the real accuracy is thus probably lower.

<sup>a</sup>To make the numbers more comparable, accuracy has been multiplied by coverage; some methods give accuracy for different degree of coverage and roughly justify this procedure. However, often it is biased toward sensitivity as specificity cannot be properly taken into account. Most features predicted with an accuracy × coverage >0.70 are of structural nature and at best only indirectly imply a certain functionality.

<sup>b</sup>Only one recent reference is given and if indicated, references therein should also be considered as other reports do not always agree with the numbers given.

translational modifications of gene products is increasing constantly, so that the complexity at the protein level is enormous. Each of these modifications may change the function of the respective gene products drastically. (The entire aspect of context-dependent gene regulation is excluded from current discussions as we are only beginning to understand the complex underlying genetic machinery. For example, promoter prediction in eukaryotes has a success of only ~35% (Table 1), and there are many other regulatory elements that we cannot predict at all.)

#### Limitations Created by Third-Party Analyses

Public releases of completely sequenced genomes exceed a rate of one per month, with thousands of function predictions therein. Gene annotation via sequence database searches is already a routine job, but even here the error rate is considerable (Table 1). The lower limit of errors in current functional annotation of large-scale sequencing projects is 8% (Brenner 1999). As errors accumulate and propagate (Bork and Bairoch 1996; Bhatia et al 1997; Smith and Zhang

1997; Bork and Koonin 1998; Pennisi 1999), it becomes more difficult to infer correct function from the many possibilities revealed by a database search. Increasing these complications is the fact that computer programs often cannot even retrieve the source of the stored information (Doerks et al. 1998).

#### Use of Complementary Information to Limit Errors in Function Prediction

Some new information can be retrieved from completely sequenced genomes, for example, function can be predicted by exploitation of genomic context.

Based on the observation that interacting proteins in one organism sometimes have homologs in other organisms fused together in a single gene, Marcotte et al. (1999a) predicted novel interactions for 50% of yeast proteins using gene fusion information. However, they noted an overlap with classical methods and an error rate of 82%. To see a signal they had to correct for domains present in many proteins (Marcotte et al. 1999a). By considering only orthologs with fission and fusion events (Enright et al. 1999, Snel et al. 2000), the signal-to-noise ratio increases and the number of predictions drops dramatically (7% of *Escherichia coli* proteins; Enright et al. 1999). With a particular question in mind, Does protein X have interaction partners?, the generation of hypotheses is extremely useful; yet to provide a general overview of protein function, it is advisable to keep the errors small. Further information can be added later, which is easier than retracting stored information. But how do we incorporate the information on error margins? Such estimates (sometimes not even the sources of the annotation) are not visible in current databases that store the results of computational approaches.

### Taking the 70% Hurdle

As noted above, most prediction schemes extrapolate from current knowledge, and many bioinformatics methods have difficulty exceeding a 70% prediction accuracy (numbers in Table 1 are often overestimates because the test sets used are usually not representative of all sequences). On one hand, current methods seem to capture important features and explain general trends; on the other hand, 30% of the features are missing or predicted wrongly. This has to be kept in mind

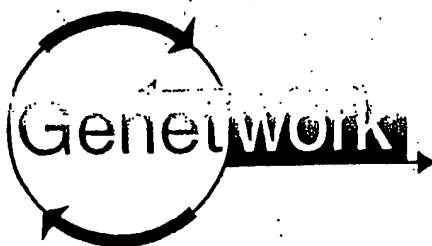
when processing the results further. Also the 70% accuracy often attaches to methods that deal with discrete objects such as sequences; making estimates about the prediction of cellular features is much more difficult as one first has to agree on semantics (or ontology in a database sense) to describe complex processes in a comparable way.

All of the above focuses on limitations in the computational prediction of qualitative features. There remains a long way to go until we are able to describe molecular processes quantitatively; current simulations of complex systems are still very rough and simplistic. However, there is still no doubt that sequence analysis is extremely powerful and that the generation of hypotheses derived by computational methods will be more and more often the first successful step in the design of experiments. If 70% of such experiments were successful, the speed of scientific discoveries would grow exponentially.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

### REFERENCES

- Anderson, L. and J. Seilhammer. *Electrophoresis* 18: 533-537.
- Andrade, M., S.I. O'Donoghue, and B. Rost. 1998. *J. Mol. Biol.* 276: 517-525.
- Bhatia, U., K. Robison, and W. Gilbert. 1997. *Science* 276: 1724-1725.
- Bork, P. and A. Bairoch. 1996. *Trends Genet.* 12: 425-427.
- Bork, P. and E.V. Koonin. 1998. *Nat. Genet.* 13: 313-318.
- Bork, P., T. Dondekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan. 1998. *J. Mol. Biol.* 283: 707-725.
- Brenner, S. 1999. *Trends Genet.* 15: 132-133.
- Buelow, K.H., M.N. Edmonson, and A.B. Cassidy. 1999. *Nat. Genet.* 21: 323-325.
- Dandekar, T. and K. Sharma. 1998. *Regulatory RNA*. Springer Verlag, Heidelberg, Germany.
- Doerks, T., A. Bairoch, and P. Bork. 1998. *Trends Genet.* 14: 248-250.
- Dunham, I., N. Shimizu, B.A. Roe, S. Chisoe, J.E. Collins, R. Bruskiewich, M. Clamp, L.J. Smink, R. Ainscough, and J.P. Almeida. 1999. *Nature* 402: 489-495.
- Ehrlich, L., M. Reczko, H. Bohr, and R.C. Wade. 1998. *Protein Eng.* 11: 11-19.
- Eisenhaber, B., P. Bork, and F. Eisenhaber. 1999. *J. Mol. Biol.* 292: 741-758.
- Enright, A.J., I. Iliopoulos, N.C. Kyrpides and C.A. Ouzounis. 1999. *Nature* 402: 86-90.
- Hanke, J., I. Zastrow, A. Aydin, G. Lehmann, S. Luft, J.G. Reich, and P. Bork. 1999. *Trends Genet.* 15: 389-390.
- Jones, D.T. 1999. *J. Mol. Biol.* 292: 195-202.
- Lupas, A. 1996. *Methods Enzymol.* 266: 513-525.
- Marcotte E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. 1999a. *Science* 285: 751-753.
- Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999b. *Nature* 402: 83-86.
- Mironov, A.A., J.W. Fickett, and M.S. Gelfand. 1999. *Genome Res.* 15: 755-771.
- Muller, A., R.M. MacCallum, and M.J.E. Sternberg. 1999. *J. Mol. Biol.* 293: 1257-1271.
- Nielsen, H., S. Brunak, and G. von Heijne. 1999. *Protein Eng.* 12: 3-9.
- Pennisi, E. 1999. *Science* 286: 447-450.
- Piatigorski, Y. and G.J. Wistow. 1991. *Science* 252: 1078-1079.
- Prestidge, D.S. 1995. *J. Mol. Biol.* 249: 923-932.
- Rost, B. 1996. *Methods Enzymol.* 266: 525-539.
- Smith, T.F. and X. Zhang. 1997. *Nat. Biotechnol.* 15: 1222-1223.
- Snel, B., P. Bork, and M. Huynen. 2000. *Trends Genet.* 16: 9-11.
- Sunyaev, S., J. Hanke, D. Brett, A. Aydin, I. Zastrow, W. Lathe, P. Bork and J. Reich. 2000. *Adv. Protein Chem.* 54: (in press).
- Teichmann, S., C. Chothia, and M. Gerstein. 1999. *Curr. Opin. Struct. Biol.* 9: 390-399.
- Tusnady, G.E. and I. Simon. 1998. *J. Mol. Biol.* 283: 489-506.



## Protein annotation: detective work for function prediction

Computer analysis of genome sequences is currently one of the essential steps for obtaining functional and structural information about the respective gene products. Database searches are used to transfer functional features from annotated proteins to the query sequences. With the increasing amount of data, more and more software robots perform this task<sup>1</sup>. While robots are the only solution to cope with the flood of data, they are also dangerous because they can currently introduce and propagate mis-annotations<sup>2,3</sup>. On the one hand, functional information is often only partially transferred (underprediction). For example, information is not usually extracted for each functional unit (protein domain) but just taken from the one-line description of the best database match (so multifunctionality is rarely considered). On the other hand, overpredictions are common because the highest-scoring database protein does not necessarily share the same or even similar functions.

### Definition and collection of uncharacterized protein families

To avoid unnecessary propagation of poor annotation, we have collected putative, poorly annotated proteins that are usually labeled as 'hypothetical' or just as 'ORF' (open reading frame). We operationally defined uncharacterized protein families (UPFs) to be families of proteins that: (1) contain members in at least three taxonomically distinct (and phylogenetically 'distant') species; and (2) do not contain (to the best of our knowledge) biochemically characterized proteins.

A collection and classification of these proteins should allow: (a) utilization of family information and thus a more detailed characterization; (b) simplification of update procedures for the entire families if functional information becomes available for at least

one member; and (c) a careful annotation of functional features that avoids the pitfalls described above.

As the number of genome sequencing projects progress, more and more of these UPFs emerge in sequence databases. We gave high priority to families that contain members in at least two of the three major kingdoms (archaea, eubacteria, eukaryotes). The original 'family' definition was based on significant hits in the statistics provided by FASTA (Ref. 4) or gapped BLAST (Ref. 5).

### Annotation of UPFs in SWISS-PROT and PROSITE databases

A serial number has been assigned to each UPF and to each of the corresponding SWISS-PROT (Ref. 6) entries. A SWISS-PROT document file lists all the current UPFs and their members in SWISS-PROT. This document is available on the WWW (Ref. 7). In the majority of cases, PROSITE entries<sup>8</sup> have already been created to document the respective family. Whenever a member of a UPF family is biochemically characterized, that family ceases to be considered as a UPF and is deleted from the list. However, information is provided that allows its history to be traced. For example:

Family: UPF0002 (DELETED)

Taxonomic range: Eubacteria

Comments: Now characterized as a family of pseudouridylate synthases (EC 4.2.1.70).

Prototype: RSUA\_ECOLI (Accession No. P33918)

PROSITE entry: PDOC00885

### Function prediction for the UPFs

The annotation is handled rather conservatively (see below) because functional overpredictions are most dangerous given the many opportunities for error propagation in sequence database<sup>2,3</sup>. Nevertheless, we intended to retrieve as many functional features as possible for each UPF using comparative analysis. Thus, each UPF was subjected to a variety of sequence analysis methods<sup>9</sup>. In brief, several members of each UPF were compared with a database of non-identical protein sequences, daily updated at the EMBL using PSI-BLAST (Ref. 5) with a conservative expected ratio of false positives ( $E = 0.001$ ) as a threshold for each iteration. Sequences were pre-processed by filtering for transmembrane<sup>10</sup> and coiled-coil regions<sup>11</sup>. A multiple alignment was constructed for each UPF using ClustalX (Ref. 12). If PSI-BLAST did not identify a relationship to characterized proteins, other iterative methods such as Wisetools (Ref. 13) and Most (Ref. 14) were applied. They also use family information, that is, give more weight to conserved positions and so on, but have the advantage that the underlying multiple alignments can be checked and improved manually (on the cost of speed and the 'easy to use' feature).

Finally, all searches were repeated using a sequence database that only contained

sequences from entirely sequenced genomes to reduce noise effects<sup>9,15</sup>. For example, PSI-BLAST E-values depend on the database size and a search might be significant using a small database but becomes insignificant if more background noise (unrelated or redundant sequences) is added.

In many cases, the iterations revealed the relationship of the UPFs with other proteins, families or superfamilies. As the main focus here was to assign functional features, the iterations have not been continued when a reasonable prediction could be made. Criteria for the latter were matches to known active site patterns or conserved motifs resembling those in PROSITE as well as the positioning of UPF members within phylogenetic trees. Transmembrane regions were identified in 13 (22%) of the 58 UPFs, although functional predictions for these 13 have not been made. Of the remaining 45 UPFs, 25 could be related to proteins with annotated functional features (Table 1).

### Pitfalls in function assignments

The predictions required careful inspection of the functional annotations of the matched database proteins. To illustrate the difficulties, Table 2 shows the result of a Blast search for UPF0002 that includes quite a few proteins with annotations (in addition to the first hits that are labeled as 'hypothetical'). Only one can give a clue about functional features; others are simply wrong, misleading or uninformative.

Another typical assignment error is caused by the sequence similarity of the query to a region that is independent from the one that was the basis for the annotation. For example, the hypothetical protein HI0722 (Accession No. P44842, ID: YIGZ\_HAEIN), a member of the UPF0029 family, shows significant similarity to two proteins (GenBank entries gil2314657 and gil2688341) in *Helicobacter pylori* and *Borrelia burgdorferi*, respectively, which are wrongly annotated as proline dipeptidases (pepQ). The annotation is based on the N-terminal homology of these two proteins with the C-terminal region of proline dipeptidase (pepQ) (gil42358) of *E. coli*, which does not harbor the catalytic activity of this enzyme.

There were even examples in which homologs scored best in PSI-BLAST (Ref. 5) that did not have the same catalytic activity because active site residues of the characterized family were not conserved. However, there were significantly lower scoring homologs with perfect matches of their (distinct) catalytic site residues to the query. For example, the UPF0046 family has clear amino acid similarity to proteases that are easily found by PSI-BLAST (Ref. 5) in the fourth iteration; yet, residues involved in metal-binding are only shared with a purple acid phosphatase family that is only picked up in the ninth iteration. The E-value of  $1e-5$  compared with proteases (E-value of  $5e-78$ ) remain considerably higher in sub-sequence iterations. Such instances have

implications for current function prediction programs in which the function of the best hit is transferred. Clearly, another generation of methods is required that include checks for the presence of functionally important residues.

### Use of phylogenetic trees

As most of the database proteins with functional annotations were only distantly related to members of the UPFs, transfer of functional information is extremely difficult and arbitrary. The majority of UPFs turned out to be related to enzymes, and based on the conservation of the active site residues one can assume that at least the basic catalytic mechanism remains the same. This, however, is of little predictive value as some families, e.g. those with the  $\alpha/\beta$  hydrolase fold collected in SCOP (Ref. 16) are huge and harbor numerous distinct catalytic activities, such as lipases, esterases, dehalogenases, peptidases, peroxidases and lyases. We have therefore constructed phylogenetic trees of selected members of the UPFs and of related, but distinct families that have been identified during the analysis (Fig. 1). On some occasions, the UPF members clearly clustered with proteins that all performed the same function (Fig. 1a), but in most of the cases the UPFs were of equal distance to distinct enzymatic activities (Fig. 1b), thus not allowing any detailed predictions.

Although the studied protein families were bound to be difficult for function predictions because a considerable number of teams were unable to find functional

TABLE 1. Predicted functional features for 25 UPFs

UPF No.	Family size <sup>a</sup>	Predicted function
02	70	Pseudouridylyl synthase
04	60	Methyltransferase
07	15	Cytidyltransferase <sup>b</sup>
08	30	ATPase
09	40	GTPase
10	10	Aldose 1-epimerase
11	10	Methyltransferase <sup>b</sup>
12	25	Nitrilase
17	30	Hydrolase
19	15	Phosphate-binding protein (TIM BARREL)
20	40	N6-adenine-specific methylase
21	50	ATPase
26	30	Two domain protein : iron/sulfur binding and amidotransferase
30	10	Amidotransferase
31	30	Sugar kinase
34	20	Pyrimidin-binding oxidoreductase (TIM BARREL)
35	20	Mutator mutt protein (7,8-dihydro-8-oxoguaninetriphosphatase)
36	70	Hydrolase
37	10	Oxydoreductase
38	35	ATPase <sup>b</sup>
42	10	ATPase
46	15	Phosphatase
49	50	N6-adenine-specific methylase
53	40	CBS domain protein
55	10	Glutaredoxin

<sup>a</sup>The numbers of family members are approximate because of daily changes in databases and loose family definitions.

<sup>b</sup>*E. coli* member also predicted by Koonin *et al.*<sup>17</sup> (UPF0007: nucleotidyltransferase).  
Abbreviation: UPFs, uncharacterized protein families.

TABLE 2. Misleading annotations: PSI-BLAST results for the UPF0002 family (first iteration)

Ranking	Annotation	Probability	Commentary
1	Gnl PID e1332795 (Z98268) hypothetical protein MTC125.33 (Mycobacterium tuberculosis)...	(2e-75)	
4	Sp P33643 SEHB_ECOLI SEHB PROTEIN	(1e-67)	SEHB is a gene name (suppressor of the temperature-sensitivity of <i>ftsH1</i> mutation) and does not give much functional insight
5	Gnl PID e1185138 (Z99112) alternative gene name: <i>ylmL</i> ; similar to hypothetical proteins [Bacillus subtilis]...	(3e-65)	
37	Sp Q12362 RIB2_YEAST DRAP DEAMINASE >gil1078332 pir IS50972 RIB2 protein - yeast (Saccharomyces cerevisiae) >gil642221 (Z21618) DRAP deaminase [Saccharomyces cerevisiae] >gil1419887 gnl PID e252279 (Z74808) ORF YOL066c [Saccharomyces cerevisiae]...	(7e-50)	The homology is not in the catalytic region and does not hold for other deaminases
40	Sp P33918 RSUA_ECOLI 16S PSEUDOURIDYLATE 516 SYNTHASE (16S PSEUDOURIDINE 516 SYNTHASE) (URACIL HYDROLYASE)	(2e-48)	Function prediction based on this protein
41	Sp Q47417 YQCB_ERWCA EXOENZYME REGULATION REGULON ORF1 >gil628643 pir IS45107 hypothetical protein 1 - Erwinia carotovora >gil496598 (X79474) ORF1 [Erwinia carotovora]...	(7e-48)	Misleading annotation, operon architecture is not conserved between species

Annotations that hamper functional predictions illustrated by the example of the UPF0002 family. Based on the recent experimental characterization of pseudouridylyl synthase<sup>18</sup>, this family has been deleted from the UPF list (see text). Nevertheless, the various, partly contradictory annotations (bold) are extremely difficult to parse for automatic function prediction programs. For brevity, the PSI-BLAST results have been cut (...).

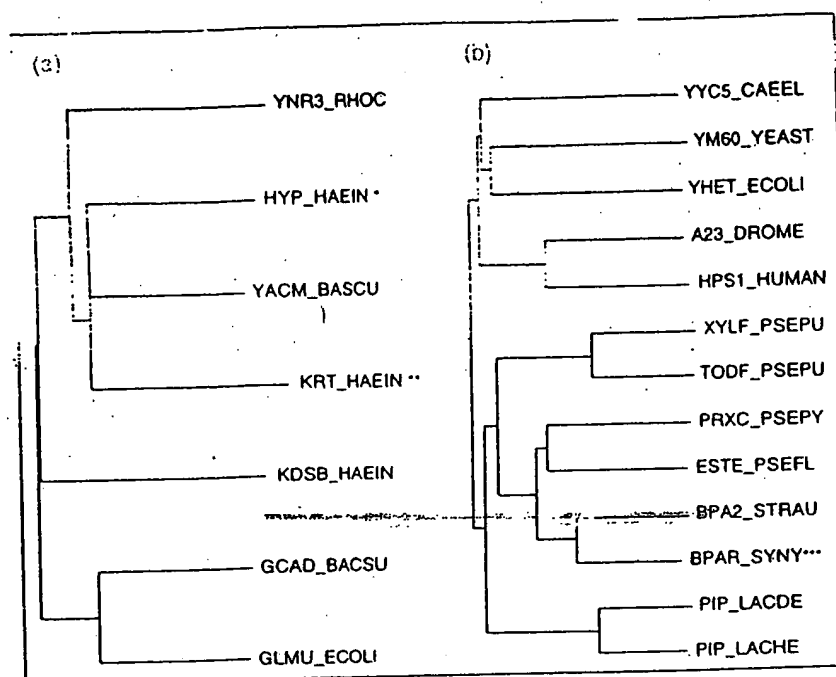


FIGURE 1. (a) Phylogenetic trees of selected members of UPF0007 that indicate a likely function as UPF0007 members with cytidyltransferase activities (red) and related uridyltransferases (blue) are more divergent (\*pir database entry, pirlg64156; \*\*pir database entry, pirls49238). (b) No clear enzymatic activity can be predicted for UPF0017 members: They clearly have the hydrolase fold but have equal distance to peroxidases (red), esterases (green), peptidases (blue) and other hydrolases (pink) (\*\*GenBank entry gii1001804). The trees were calculated using CLUSTALX (Ref. 12).

features therein, it is noteworthy that there was not a single case in which we were able to predict the precise mechanism and the substrate specificity. Nevertheless, the information about an enzymatic activity and the likely reaction mechanisms of the 25 UPFs should prove useful for the analysis of upcoming genome sequences.

## Annotation with the right level of precision helps in future projects

In summary, we were able to provide some functional annotation for more than 700 of about 1300 proteins clustered in 25 of the 58 distinct UPFs. Most of them are currently named 'hypothetical protein' so that their annotation adds enormous value to these sequences. For another 13 UPFs currently containing about 250 proteins, the presence of transmembrane regions was recorded. This annotation is now being incorporated into PROSITE and SWISS-PROT so that these features can be assigned to newly sequenced genes as well.

The difficulties we faced in assigning functions by sequence similarity also indicate that many of the automatic predictions by most of the software robots are probably erroneous. Because of the current policies of most of the sequence databases, correction of annotations is very hard to realize. Thus, there should be a combined effort by the database teams, the authors of the current entries, and the community, to work towards a careful functional annotation of all the

## References

- 1 Boguski, M. and McEntyre, J. (1994) *Trends Biochem. Sci.* 19, 71
- 2 Bork, P. and Bairoch, A. (1996) *Trends Genet.* 12, 425-427
- 3 Bhatia, U., Robinson, K. and Gilbert, W. (1997) *Science* 276, 1724-1725
- 4 Pearson, W.R. and Miller, W. (1992) *Methods Enzymol.* 210, 575-601
- 5 Altschul, S.F. et al. (1997) *Nucleic Acids Res.* 25, 3389-3402
- 6 Bairoch, A. and Apweiler, R. (1998)

*Nucleic Acids Res.* 26, 38-42

7 <http://www.expasy.ch/cgi-bin/lists/uplist.txt>

8 Bairoch, A., Bucher, P. and Hofmann, K. (1998) *Nucleic Acids Res.* 25, 217-221

9 Bork, P. and Gibson, T. (1996) *Methods Enzymol.* 266, 162-184

10 Von Heijne, G. (1992) *J. Mol. Biol.* 225, 487-494

11 Lupas, A., van Dyke, M. and Stock, J. (1991) *Science* 252, 1162-1164

12 Thompson, J.D. et al. (1997) *Nucleic Acids Res.* 25, 4876-4882

13 Birney, E., Thompson, J.D. and Gibson, T.J. (1996) *Nucleic Acids Res.* 24, 2730-2739

14 Tatusov, R.L., Altschul, S.F. and Koonin, E.V. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 91, 12091-12095

15 Bork, P. and Koonin, E.V. (1998) *Nat. Genet.* 18, 313-318

16 Murzin, A.G., Brenner, S.E., Hubbard, T. and Chotia, C. (1995) *Mol. Biol.* 247, 536-540

17 Koonin, E., Mushegian, A., Galperin, M. and Walker, D. (1997) *Mol. Microbiol.* 25, 619-637

18 Wrzesinski, J. et al. (1995) *Biochemistry* 34, 8904-8913

Tobias Doerks

doerks@embl-heidelberg.de

EMBL Meyerhofstrasse 1, 69012 Heidelberg  
and Max-Planck-Center for Molecular  
Medicine, Berlin-Buch, Germany.

Amos Bairoch

amos.bairoch@medecine.unige.ch

Swiss Bioinformatics Institute and  
University of Geneva, Switzerland.

Peer Bork

bork@embl-heidelberg.de

## TECHNICAL TIPS ONLINE

<http://www.elsevier.com/locate/tto> ♦ <http://www.elsevier.nl/locate/tto>

Editor Adrian Bird

Institute for Cell and Molecular Biology at the University of Edinburgh

Protocols are now featured in *Technical Tips Online*, in addition to peer-reviewed *Technical Tips* articles (novel applications or significant improvements on existing methods). Protocols incorporate all the features that are currently available in *Technical Tips* articles: comment facility; links to Medline abstracts; product information and so on.

New Core Protocol articles published recently in *Technical Tips Online* include:

- Mitchell, T.J. and Morely, B.J. (1998) Isolation of RNA and analysis by northern blotting and primer extension *Technical Tips Online* (<http://www.elsevier.com/locate/tto>) P01236

ities of IL-8 and  
f IL-8 is required  
appears critical  
ch packs against  
mains of CXCR1  
with IL-8 and to  
28-31 We suggest;  
ctly with the N-  
is supported by  
terminal residues  
veral residues in

m the important  
condary binding  
re available data  
secondary N-loop  
ain and that the  
Arg-199, Arg-  
esis in CXCR1

Wood for assistance  
GN and INSIGHT-

rna, D. P. Cerretti,  
Biol. Chem. 268,

ruk, J. Biol. Chem.

343 (1994).

ane, J. Biol. Chem.

, FEBS Lett. 338,

## [5] Alanine Scan Mutagenesis of Chemokines

By JOSEPH HESSELGESSER and RICHARD HORUK

### Introduction

Chemokines are a family of chemotactic cytokines that play a critical role in the regulation and trafficking of immune cells.<sup>1</sup> Some chemokines have also been shown to be suppressive factors<sup>2</sup> of human immunodeficiency virus type-1 (HIV-1). There are currently three families of chemokines, grouped according to the arrangement of their invariant cysteines. The C-X-C branch in which the first two cysteines are separated by an intervening amino acid includes interleukin-8 (IL-8), melanoma growth stimulating activity (MGSA), platelet factor 4 (PF4), and stromal cell derived factor-1 (SDF-1).<sup>3</sup> This group is further divided by the presence (IL-8, MGSA) or absence (SDF-1, PF4) of an E-L-R motif prior to the first N-terminal cysteine. The C-X-C chemokines are generally chemoattractors and activators of neutrophils, exceptions being PF4, which also attracts fibroblasts and monocytes and SDF-1, which is a chemoattractant for T cells. The C-C chemokines are chemoattractants and cellular activators for monocytes, basophils, eosinophils, and lymphocytes. This is the largest family of chemokines and includes the macrophage inflammatory proteins MIP-1 $\alpha$  and MIP-1 $\beta$ , RANTES, monocyte chemoattractant proteins MCP-1 to MCP-4, I-309, and eotaxin.<sup>3</sup> The third subdivision of the chemokines is the C branch, which is characterized by its only member lymphotactin, in which only two cysteines are conserved.<sup>4</sup>

In addition to their normal role in immune cell function, chemokines play an important part in a number of autoimmune diseases such as multiple sclerosis and rheumatoid arthritis,<sup>3</sup> and chemokine receptors are gateways

<sup>1</sup> T. Schall, in "The Cytokine Handbook" (A. Thompson, ed.), p. 419. Academic Press, San Diego, 1994.

<sup>2</sup> F. Cocchi, A. L. DeVico, A. Garzino-Demo, S. K. Arya, R. C. Gallo, and P. Lusso, *Science* 270, 1811 (1995).

<sup>3</sup> R. Horuk, ed., "Chemoattractant Ligands and Their Receptors." CRC Press, Boca Raton, Florida, 1996.

<sup>4</sup> G. S. Kelner, J. Kennedy, K. B. Bacon, S. Kleyensteuber, D. A. Largaespada, N. A. Jenkins, N. G. Copeland, J. F. Bazan, K. W. Moore, T. J. Schall, and A. Zlotnik, *Science* 266, 1395 (1994).

of infection for HIV-1 and the malaria parasite.<sup>5-8</sup> Thus, it is essential to understand how chemokines interact with their cellular receptors to produce their biological effects to develop therapeutic approaches of intervention.

### Rationale

Structure-function relationships between ligands and their receptors can be examined in a number of ways. Site-directed mutagenesis, particularly alanine scan mutagenesis, has been a successful technique for identifying functionally important residues in proteins. This approach has been applied to a number of receptors and ligands as exemplified for human growth hormone (hGH), IL-8, MGSA, CXCR1 (IL-8RA), and CXCR2 (IL-8RB).<sup>9-13</sup> In a variation of alanine scan mutagenesis the charged residues of a molecule only are replaced, and thus residues forming ion pairs with charged residues in a cognate receptor can be identified. In addition, substitutions of charged amino acid residues by residues with increasing side-chain length or bulkiness can be used to determine what can be tolerated for receptor binding. Finally, replacement of hydrophobic residues by alanine gives a glimpse of important nonaqueous interactions between ligand-receptor pairs.

Site-directed mutagenesis to determine structure-function relationships for ligand-receptor pairs is preferred over ligand truncation mutants because it minimizes structural modifications to the native molecule. Another advantage of site-directed mutagenesis is that, if structural data exist, the surface residues of a protein can be substituted in a predictable pattern, one at a time or in groups. The substitution of groups of amino acids

<sup>5</sup> H. Choe, M. Farzan, Y. Sun, N. Sullivan, B. Rollins, P. D. Ponath, L. J. Wu, C. R. Mackay, G. LaRosa, W. Newman, N. Gerard, C. Gerard, and J. Sodroski, *Cell (Cambridge, Mass.)* 85, 1135 (1996).

<sup>6</sup> T. Dragic, V. Litwin, G. P. Allaway, S. R. Martin, Y. X. Huang, K. A. Nagashima, C. Cayanan, P. J. Maddon, R. A. Koup, J. P. Moore, and W. A. Paxton, *Nature (London)* 381, 667 (1996).

<sup>7</sup> B. J. Doranz, J. Rucker, Y. J. Yi, R. J. Smyth, M. Samson, S. C. Peiper, M. Parmentier, R. G. Collman, and R. W. Doms, *Cell (Cambridge, Mass.)* 85, 1149 (1996).

<sup>8</sup> R. Horuk, *Immunol. Today* 15, 169 (1994).

<sup>9</sup> B. C. Cunningham and J. A. Wells, *Science* 244, 1081 (1989).

<sup>10</sup> C. A. Hébert, R. V. Vitangcol, and J. B. Baker, *J. Biol. Chem.* 266, 18989 (1991).

<sup>11</sup> J. Hesselgesser, C. Chitnis, L. Miller, D. J. Yansura, L. Simmons, W. Fairbrother, C. Kotts, C. Wirth, B. Gillette-Castro, and R. Horuk, *J. Biol. Chem.* 270, 11472 (1995).

<sup>12</sup> C. A. Hébert, A. Chuntharapai, M. Smith, T. Colby, J. Kim, and R. Horuk, *J. Biol. Chem.* 268, 18549 (1993).

<sup>13</sup> S. R. Leong, R. C. Kabakoff, and C. A. Hébert, *J. Biol. Chem.* 269, 19343 (1994).

[5]

from IL-8 with exhibiting high-low affinity) prior in helping to its specificity.<sup>14</sup> A is BB-10010, a at position 27,<sup>15</sup> but does not fit the wild-type c situations in w high concentrat that could redu

### Protocols

#### Expression Sys

Recombinant methods for p described incl Jolla, CA). The *coli* strain (Stra of *lacUV5*. This RNA polymer: of isopropylthi prior to the cl heat-stable ent of a soluble se

Once an a kine gene of ir ate a variety c

<sup>14</sup> H. B. Lowman, and W. J. Fair

<sup>15</sup> M. G. Hunter, Dexter, A. H. J

<sup>16</sup> I. Lindley, H. / P. Peveri, B. D U.S.A. 85, 9195

<sup>17</sup> R. Horuk, D. Unemori, J. Bi

<sup>18</sup> M. C. N. Johns

<sup>19</sup> J. Zagorski and



essential to  
tensors to pro-  
cesses of inter-

ir receptors  
sis, particu-  
e for identi-  
ch has been  
for human  
CXCR2 (IL-  
1 residues of  
pairs with  
ition, substi-  
easing side-  
be tolerated  
dues by ala-  
een ligand-

relationships  
mutants be-  
le. Another  
ta exist, the  
ble pattern,  
amino acids

, C. R. Mackay,  
nbridge, Mass.)

Nagashima, C.  
ature (London)

M. Parmentier,  
).

(1991).  
rother, C. Koits,  
95).  
s, J. Biol. Chem.

3 (1994).

from IL-8 with those from MGSA to produce a hybrid protein capable of exhibiting high-affinity binding to CXCR1 (native MGSA binds only with low affinity) provides an excellent example of the utility of this technique in helping to identify domains of IL-8 that are responsible for receptor specificity.<sup>14</sup> A further example of the application of alanine replacement is BB-10010, a variant of MIP-1 $\alpha$ , that has a single substitution Asp to Ala at position 27.<sup>15</sup> This substitution results in a molecule that is fully active but does not form the extremely high molecular weight multimers that the wild-type chemokine does. BB-10010 has thus found favor in clinical situations in which patients undergoing chemotherapy can be dosed with high concentrations of the chemokine without fear of aggregate formation that could reduce its efficacy.

## Protocols

### Expression Systems

Recombinant expression in *Escherichia coli* is one of the most common methods for producing chemokines.<sup>16-19</sup> A number of vectors have been described including the pET-3 and PET-11 series, from Stratagene (La Jolla, CA). These are generally used in combination with a BL21(DE3) *E. coli* strain (Stratagene) that contains T7 RNA polymerase under the control of *lacUV5*. This expression system allows genes placed downstream of the RNA polymerase binding site (in pET vectors) to be expressed on addition of isopropylthiogalactoside (IPTG). Upstream cloning of a signal sequence, prior to the chemokine gene, such as human growth hormone (hGH) or heat-stable enterotoxin II (hstII)<sup>11</sup> has been shown to aid in the expression of a soluble secreted protein.

Once an appropriate vector construct has been made with the chemokine gene of interest, site-directed mutagenesis can be performed to generate a variety of clones. The site-directed alanine substitution mutants can

<sup>14</sup> H. B. Lowman, P. H. Slag, L. E. DeForge, C. M. Wirth, B. L. Gillette-Castro, J. H. Bourell, and W. J. Fairbrother, *J. Biol. Chem.* 271, 14344 (1996).

<sup>15</sup> M. G. Hunter, L. Bawden, D. Brotherton, S. Craig, S. Cribbes, L. G. Czaplewski, T. M. Dexter, A. H. Drummond, A. H. Gearing, and C. M. Heyworth, *Blood* 86, 4400 (1995).

<sup>16</sup> I. Lindley, H. Aschauer, J.-M. Seifert, C. Lam, W. Brunowsky, E. Kownatzki, M. Thelen, P. Peveri, B. Dewald, V. von-Tscharnier, A. Walz, and M. Baggiolini, *Proc. Natl. Acad. Sci. U.S.A.* 85, 9199 (1988).

<sup>17</sup> R. Horuk, D. G. Yansura, D. Reilly, S. Spencer, J. Bourell, W. Henzel, G. Rice, and E. Unemori, *J. Biol. Chem.* 268, 541 (1993).

<sup>18</sup> M. C. N. Johnson et al., *J. Biol. Chem.* 271, 10853 (1996).

<sup>19</sup> J. Zagorski and J. E. DeLarco, *Protein Expression Purif.* 5, 337 (1994).

be made by constructing oligonucleotide probes that cover the desired mutation followed by polymerase chain reaction (PCR) to fill in the rest of the chemokine gene sequence. There are a variety of kits available from several companies that use similar, simple, and easy-to-follow methods to insert the desired mutations. These kits include the Transformer Site-Directed Mutagenesis Kit (Clontech, Palo Alto, CA) and the Chameleon Double-Stranded Site-Directed Mutagenesis Kit (Stratagene). The basic principle behind these kits is illustrated in Fig. 1. With these bacterial expression systems it is simple to produce small amounts of proteins using 1-liter shake flasks; alternatively, this process can be scaled up to larger 10- to 20-liter fermentation runs to produce gram quantities of protein. In general, chemokines can be produced as secreted, fully folded proteins,<sup>11,17</sup> but sometimes problems with folding and also with the generation of N-terminal addition mutants can occur (see Horuk *et al.*<sup>19a</sup> for a discussion).

Chemokines have also been produced in mammalian expression systems.<sup>20</sup> A variety of cell lines, such as human kidney 293 cells, chinese hamster ovary (CHO) cells, and COS cells, can be used. Mammalian expression vectors are generally larger than *E. coli* vectors, with viral promoters and enhancers (pSI, pCI, and pCI-neo from Promega, Madison, WI; pcDNA3 from Invitrogen, San Diego, CA; or pPUR from Clontech). The use of an upstream signal sequence, such as hGH, CD4, or human immunoglobulin (Ig), can greatly enhance the production of a soluble secreted protein. A major advantage of mammalian expression is that the secreted protein is usually correctly folded. In addition with the introduction of defined, nonserum-containing medium (HyQ-CCM 1 to 5, Hyclone, Logan, UT, and CHO-S-SFM II, GIBCO-BRL, Grand Island, NY) downstream purification steps are easier, and there is less risk of contamination with lipopolysaccharide (LPS), which is frequently a problem with *E. coli* expression systems.

### Peptide Synthesis

An alternative to protein expression that has found increasing use for the generation of chemokine mutants is peptide synthesis. Although this can be an expensive process (\$20/amino acid plus purification), it is quick compared to expression methods. For a discussion of peptide synthesis of chemokines see Clark-Lewis *et al.*<sup>20a</sup>

<sup>19a</sup> R. Horuk, D. Reilly, and D. Yansura, *Methods Enzymol.* 287 [1], 1997 (this volume).

<sup>20</sup> E. Balentien, J. H. Han, H. G. Thomas, D. Z. Wen, A. K. Samantha, C. O. Zachariae, P. R. Griffin, R. Brachmann, W. L. Wong, K. Matsushima, and R. Derynck, *Biochemistry* 29, 10225 (1990).

<sup>20a</sup> I. Clark-Lewis, L. Vo, P. Owen, and J. Anderson, *Methods Enzymol.* 287 [16], 1997 (this volume).

FIG. 1. Strategy for plasmid DNA. An restriction site and the using dNTPs and T7 with restriction enzyme plasmid DNA. Trax (Stratagene)] that are (Plasmid Mini Kit, C enzyme. Transfect in the correct sequence

the desired  
l in the rest  
ailable from  
ow methods  
former Site-  
Chameleon  
) . The basic  
se bacterial  
roteins using  
up to larger  
f protein. In  
proteins,<sup>11,17</sup>  
eneration of  
discussion).  
pression sys-  
ells, chinese  
alian expres-  
l promoters  
adison, WI;  
ontech). The  
an immuno-  
ble secreted  
the secreted  
roduction of  
lone, Logan,  
downstream  
ination with  
. coli expres-

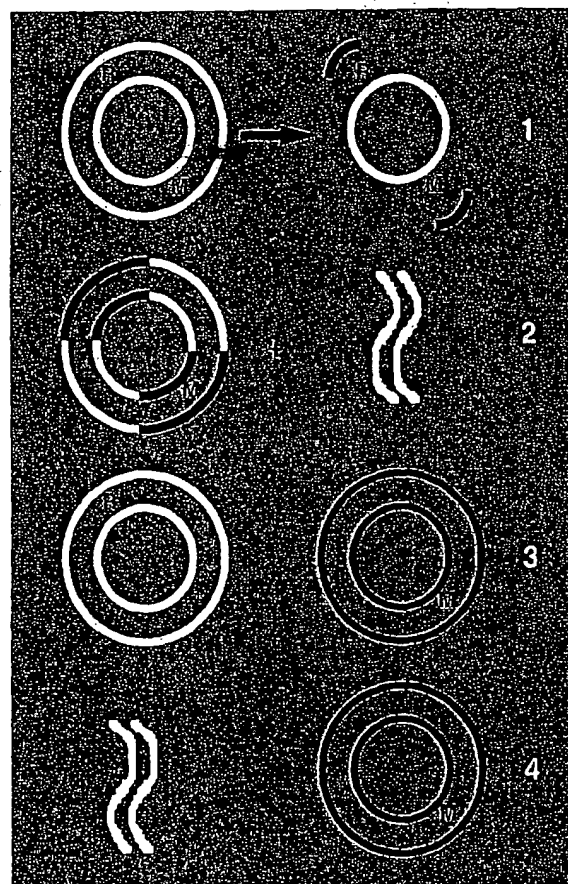


FIG. 1. Strategy for generation of alanine scan mutants. (1) Denature the double-stranded plasmid DNA. Anneal the two oligonucleotide primers, the R primer altering a unique restriction site and the M primer containing the mutation codon. (2) Perform primer extension using dNTPs and T7 or T4 DNA polymerase and ligation with T4 DNA ligase. Digest DNA with restriction enzyme for unique site. Renature DNA to result in a mixture of linear and plasmid DNA. Transform into *E. coli* strains [BMH 71-18 mutS (Clontech) or XLmutS (Stratagene)] that are unable to perform DNA mismatch repair. (3) Recover plasmid DNA (Plasmid Mini Kit, Qiagen, Chatsworth, CA). (4) Perform second digestion with restriction enzyme. Transfect into *E. coli* and recover final plasmid. Sequence the mutated gene to verify the correct sequence.

using use for  
lthough this  
) , it is quick  
synthesis of

is volume).  
O. Zachariae,  
k, *Biochemistry*

### Purification

Most chemokines are very basic proteins with isoelectric points between pH 8.0 and 10.5. In contrast most *E. coli* and serum proteins from mammalian cell expression have acidic isoelectric points. Thus chemokines can be purified at neutral pH by a combination of S-Sepharose, heparin-Sepharose, and hydrophobic interaction (phenyl-Sepharose) chromatography (see Horuk *et al.*<sup>19a</sup>). This procedure coupled with reversed-phase high-performance liquid chromatography (HPLC) can give chemokines that are >95% pure based on silver staining sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and amino acid composition.<sup>11</sup> Soluble chemokine supernatants from *E. coli* or mammalian cell supernatants can be purified by this method. Alternatively, the construction of expression plasmids containing nucleotides encoding hexahistidine, FLAG peptide, glutathionine, or immunoglobulin sequences engineered into the cDNAs of specific chemokines can greatly aid in their purification.<sup>21-23</sup> The protein epitope can be engineered into the amino- or carboxyl-terminal region of the molecule (the position is largely dependent on the maintenance of full biological activity).

### Biological Activity

Once wild-type and mutant chemokine proteins have been generated and purified they are usually analyzed by receptor-binding and bioactivity studies to determine structure-function relationships. For receptor-binding studies, a wide variety of radiolabeled chemokines are available commercially (Dupont NEN, Boston MA; Amersham, Arlington Heights, IL) or can be iodinated.<sup>23a</sup> Binding assays can be carried out on whole cells or membranes prepared from cells that express the appropriate chemokine receptors. Scatchard analysis of the binding data can then be used to determine the relative receptor-binding affinity of the mutant chemokine, which is defined as the  $K_D$  of the wild-type/ $K_D$  mutant  $\times 100\%$ . Biological assays for chemokines are numerous and are described elsewhere.<sup>23b</sup> However, one biological assay that is used as a standard is chemotaxis. All chemokines induce a migratory response in target cells that express functional G-protein-coupled chemokine receptors. In addition, a primary response to li-

<sup>21</sup> H. M. Sassenfeld, *Trends Biotechnol.* 8, 88 (1990).

<sup>22</sup> A. S. Robeva, R. Woodard, D. R. Luthin, H. E. Taylor, and J. Linden, *Biochem. Pharmacol.* 51, 545 (1996).

<sup>23</sup> A. Kuusinen, M. Arvola, C. Oker-Blom, and K. Keinänen, *Eur. J. Biochem.* 233, 720 (1995).

<sup>23a</sup> G. L. Bennett and R. Horuk, *Methods Enzymol.* 288 [10] (1997).

<sup>23b</sup> D. Baly, U. Gibson, D. Allison, and L. DeForge, *Methods Enzymol.* 287 [6], 1997 (this volume); R. C. Newton and K. Vaddi, *Methods Enzymol.* 287 [12], 1997 (this volume).

gand-receptor stores. Calcitonin receptor-like function

### Applications

The following alanine scan: structure-function antigen receptor generating a response to invasion of host but is a very

### Plasmid Construction

Plasmid construction of vectors and with reference has been previously to aid promoter that containing the sequence as described the MGSA gene mutant the expression respective coding DNA is synthesized (Perkin-Elmer) appropriate amount tions. DNA is (Qiagen, Chatsworth, CA)

The MGSA restriction enzyme NaCl, 10 mM for 2 hr. The *EcoRI* follows. The mixture is hangs with the

<sup>23c</sup> S. R. McColl *et al.*, *Methods Enzymol.* 288 [10] (1997).

<sup>24</sup> C. N. Chang, M.

tric points between  
eins from mamma-  
chemokines can be  
eparin-Sepharose,  
omatography (see  
phase high-perfor-  
ines that are >95%  
polyacrylamide gel  
position.<sup>11</sup> Soluble  
ll supernatants can  
tion of expression  
e, FLAG peptide,  
d into the cDNAs  
n.<sup>21-23</sup> The protein  
-terminal region of  
aintenance of full

ve been generated  
ing and bioactivity  
or receptor-binding  
available commer-  
on Heights, IL) or  
on whole cells or  
opriate chemokine  
n be used to deter-  
chemokine, which  
s. Biological assays  
here.<sup>23b</sup> However,  
is. All chemokines  
functional G-pro-  
ary response to li-

1, *Biochem. Pharmacol.*

*biochem.* 233, 720 (1995).

*mol.* 287 [6], 1997 (this  
, 1997 (this volume).

gand-receptor binding is the induction of the intracellular release of calcium stores. Calcium flux assays are also popular methods of assessing chemokine function.<sup>23c</sup>

### Applications

The following section gives an outline for the specific production of alanine scan mutants of MGSA that have been generated to probe structure-function relationships of the chemokine receptors CXCR2 and duffy antigen receptor for chemokines (DARC).<sup>11</sup> This approach is useful in generating a mutant of MGSA, E6A, that is able to inhibit malaria parasite invasion of human erythrocytes by receptor blockade of the receptor DARC but is a very poor agonist of the CXCR2 receptor on neutrophils.

### Plasmid Construction

Plasmid construction and mutagenesis are achieved by a combination of vectors and kits described above (see also Fig. 1) and are illustrated with reference to the MGSA mutants. The pMG34 *E. coli* secretion vector has been previously described.<sup>17</sup> This vector contains the hstII signal sequence to aid in secretion of the chemokine and an alkaline phosphatase promoter that is induced when the *E. coli* cells are grown in a low phosphate-containing medium.<sup>24</sup> The MGSA gene sequence is fused to the hstII sequence as described.<sup>17</sup> The resulting plasmid contains an *EcoRI* site 5' to the MGSA gene and a *BsaJI* site 3' to the MGSA sequence. For each mutant the entire MGSA gene is synthesized (~219 bases) to include the respective codon changes for the amino acid substitution. Double-stranded DNA is synthesized by the use of an AmpliTaq DNA Polymerase kit (Perkin-Elmer, Roche Molecular Systems, Branchburg, NJ) with appropriate amounts of DNA and polymerase per the manufacturer's specifications. DNA is purified by the use of QIAquick Nucleotide Removal Kit (Qiagen, Chatsworth, CA) per the manufacturer's specifications.

The MGSA DNA and pMG34 vector are cut separately with *EcoRI* restriction endonuclease (New England Biolabs, Beverly, MA) in 50 mM NaCl, 10 mM Tris-HCl, 10 mM MgCl<sub>2</sub>, 1 mM dithiothreitol (DTT) at 37° for 2 hr. The temperature is then raised to 65° for 20 min to inactivate *EcoRI* followed by addition of *BsaJI* endonuclease (New England Biolabs). The mixture is then incubated at 60° for 2 hr. This results in *EcoRI* overhangs with the MGSA DNA and cleaved plasmid. The DNA mixture is

<sup>23c</sup> S. R. McColl and P. H. Naccache, *Methods Enzymol.* 288 [18] (1997); K. B. Bacon, *Methods Enzymol.* 288 [22] (1997).

<sup>24</sup> C. N. Chang, M. Rey, B. Bochner, H. Heyneker, and G. Gray, *Gene* 55, 189 (1987).

purified by the QIAquick kit as above. The mutated MGSA DNA is then ligated into the vector by the use of T4 DNA ligase (20 units, 0.3 Weiss units; New England Biolabs) in 50 mM Tris-HCl (pH 7.5), 10 mM  $MgCl_2$ , 10 mM DTT, 0.5  $\mu M$  ATP, 25  $\mu g/ml$  bovine serum albumin (BSA), 200 ng vector, and 100 ng insert at 16° for 16 hr at a final volume of 20  $\mu l$ . Ligase activity is stopped by heat inactivation at 65° for 20 min.

The plasmid is transfected into *E. coli* by the  $CaCl_2$ /heat shock method<sup>25</sup> and grown in LB broth with 50  $\mu g/ml$  carbenicillin. The plasmid is recovered from the *E. coli* by the use of Qiagen Plasmid Mini Kit as above. Finally, the plasmid is sequenced through the entire gene and restriction sites to verify the correct orientation and nucleotide sequence. Following plasmid recovery and sequencing, the plasmid is transformed into *E. coli*, cells are grown overnight in low phosphate medium, and the MGSA is purified as described.<sup>11</sup>

#### Studies with Melanoma Growth Stimulating Activity Mutants

Because MGSA binds with high affinity to the chemokine receptors CXCR2 (IL-8RB) and DARC,<sup>26,27</sup> receptor-binding studies with the MGSA mutants are carried out in cells expressing these receptors. Cells are incubated with <sup>125</sup>I-labeled MGSA in the absence and presence of unlabeled MGSA or MGSA mutants; typical competition binding curves are shown in Fig. 2. The E6A mutant of MGSA is able to bind with high affinity to DARC ( $K_D = 7$  nM compared to  $K_D = 3.5$  nM for MGSA), but in contrast binds with low affinity to CXCR2 ( $K_D = 476$  nM for E6A compared to  $K_D = 2.3$  nM for MGSA). In contrast, the R8A mutant exhibits low-affinity binding to both receptors (for CXCR2  $K_D = 850$  nM and for DARC  $K_D = 157$  nM). Thus, the Arg-8 residue of MGSA appears to be crucial for expression of high-affinity binding for both receptors.

In light of the binding data obtained for the E6A and R8A mutants, functional assays are carried out both with DARC and with CXCR2. DARC is known to be a cofactor for the invasion of human erythrocytes by the malarial parasite *Plasmodium vivax*<sup>28</sup> and the related monkey parasite *Plasmodium knowlesi*.<sup>29</sup> The C-X-C chemokines IL-8 and MGSA have

<sup>25</sup> M. Mandel and A. Higa, *J. Mol. Biol.* 53, 159 (1970).

<sup>26</sup> J. Lee, R. Horuk, G. C. Rice, G. L. Bennett, T. Camerato, and W. I. Wood, *J. Biol. Chem.* 267, 16283 (1992).

<sup>27</sup> R. Horuk, C. E. Chitnis, W. C. Darbonne, T. J. Colby, A. Rybicki, T. J. Hadley, and L. H. Miller, *Science* 261, 1182 (1993).

<sup>28</sup> L. H. Miller, S. J. Mason, D. F. Clyde, and M. H. McGinniss, *N. Engl. J. Med.* 295, 302 (1986).

<sup>29</sup> L. H. Miller, S. J. Mason, J. A. Dvorak, M. H. McGinniss, and I. K. Rothman, *Science* 189, 561 (1975).

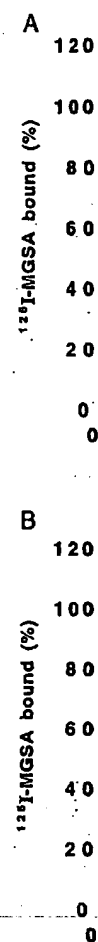


FIG. 2. MGSA cc <sup>125</sup>I-MGSA and incre (A) human erythrocy incubated 1 hr at 4° v MGSA mutants. (Ad

A DNA is then  
units, 0.3 Weiss  
10 mM  $MgCl_2$ ,  
nin (BSA), 200  
lume of 20  $\mu$ l.  
1 min.  
shock method<sup>25</sup>  
aid is recovered  
above. Finally,  
traction sites to  
lowing plasmid  
E. coli, cells are  
iSA is purified

15

ckine receptors  
with the MGSA  
Cells are incu-  
e of unlabeled  
rves are shown  
high affinity to  
but in contrast  
A compared to  
its low-affinity  
nd for DARC  
s to be crucial

R8A mutants,  
XCR2. DARC  
rocytes by the  
onkey parasite  
1 MGSA have

od, *J. Biol. Chem.*

Hadley, and L. H.

ed 295,302 (1986).  
hman, *Science* 189,

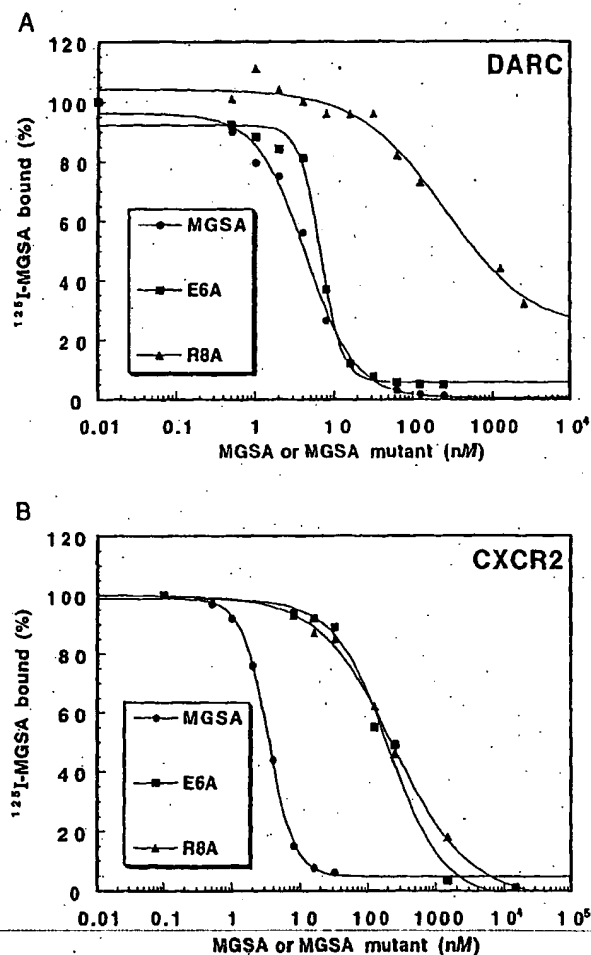


FIG. 2. MGSA competition binding studies. Competition binding was studied between  $^{125}I$ -MGSA and increasing concentrations of unlabeled MGSA mutants E6A and R8A for (A) human erythrocytes and (B) human kidney cells transfected with CXCR2. Cells were incubated 1 hr at 4° with  $^{125}I$ -MGSA in the presence of increasing amounts of the unlabeled MGSA mutants. (Adapted from Ref. 11 with permission.)

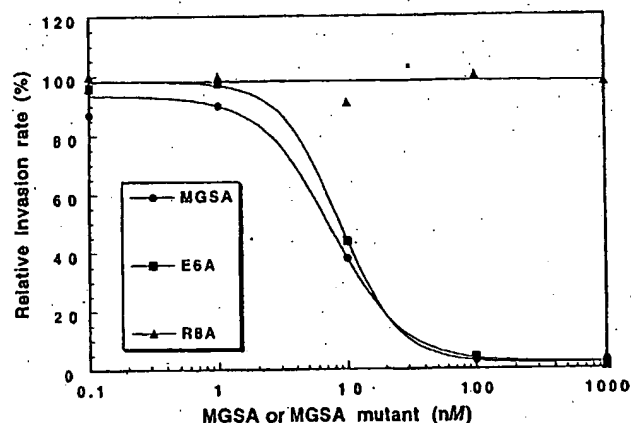


FIG. 3. Inhibition of erythrocyte invasion by *P. knowlesi* by MGSA and MGSA mutants. The invasion rates are expressed as a percentage of the rate of invasion in the absence of chemokines. Inhibition of invasion  $EC_{50}$  values are as follows: MGSA, 7 nM; E6A, 8.6 nM; R8A,  $>1 \mu M$ . (Adapted from Ref. 11 with permission.)

been shown to dose-responsively inhibit both parasite binding and invasion.<sup>27</sup> Thus, the ability of MGSA and the MGSA mutants E6A and R8A to block *P. knowlesi* invasion of human erythrocytes is assessed as previously described.<sup>11</sup> As shown in Fig. 3 the E6A MGSA mutant is able to block

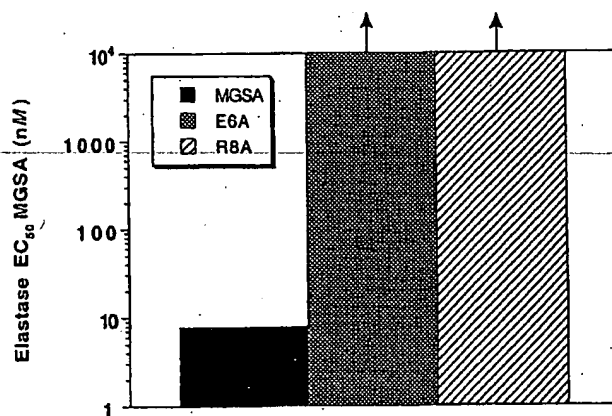


FIG. 4. MGSA and MGSA mutant stimulation of elastase release from human neutrophils. The  $EC_{50}$  value is defined as the concentration of MGSA or mutant required for half-maximal release of elastase from neutrophils. (Adapted from Ref. 11 with permission.)

malarial parasite MGSA (for MGSA line with the chemokine) is unable to block

The mutants express CXCR2 release assay (for are biologically 8 nM). Thus, it us to identify a human erythrocyte be useful therapeutic rocyte invasion

## Conclusion

The example illustrates the specific residues but also for general chemokine receptor diseases, the use for binding domain would be useful

## Introduction

Chemokines attacks by invading is selective for a the assays available

<sup>1</sup> T. Schall, in "The Diego, 1994.



[5]



d MGSA mutants.  
in the absence of  
1M; E6A, 8.6 nM;

ding and inva-  
5A and R8A to  
d as previously  
s able to block



human neutrophils.  
d for half-maximal  
ion.)

[6]

BIOLOGICAL ASSAYS FOR C-X-C CHEMOKINES

69

malarial parasite invasion in a dose-responsive manner similar to wild-type MGSA (for MGSA  $EC_{50} = 7$  nM, for E6A  $EC_{50} = 8.6$  nM). However, in line with the change in binding affinity of R8A to DARC, the R8A mutant is unable to block malarial invasion at concentrations up to 1  $\mu$ M.

The mutants are then tested for biological activity on neutrophils, which express CXCR2, using elastase release assays. The data from the elastase release assay (Fig. 4) clearly show that the MGSA mutants E6A and R8A are biologically inactive ( $EC_{50}$  10  $\mu$ M) compared to wild-type MGSA ( $EC_{50}$  8 nM). Thus, the use of alanine scan mutagenesis of MGSA has allowed us to identify an MGSA mutant, E6A, that can block malarial invasion of human erythrocytes but will not activate neutrophils, properties that may be useful therapeutically in the design of small molecules that inhibit erythrocyte invasion by *P. vivax* but have no effect on neutrophils.

### Conclusion

The example of alanine scan mutagenesis described for MGSA clearly illustrates the powerful nature of this technique not only for identifying specific residues involved in the binding of ligands to their native receptors but also for generating potentially useful drugs. Because chemokines and chemokine receptors play an important role in a variety of acute and chronic diseases, the use of alanine scanning mutagenesis to accurately define receptor binding domains could aid in the design of molecular antagonists that would be useful therapeutically.

## [6] Biological Assays for C-X-C Chemokines

By DEBORAH BALLY, URSULA GIBSON, DAVID ALLISON,  
and LAURA DEFORGE

### Introduction

Chemokines play a major role in mobilizing leukocytes to ward off attacks by invading pathogens.<sup>1</sup> Each of the two major classes of chemokines is selective for a particular group of immune cells. In this chapter, we discuss the assays available to measure the biological activities of one of these

<sup>1</sup> T. Schall, in "The Cytokine Handbook" (A. Thompson, ed.), p. 419. Academic Press, San Diego, 1994.

## Review

# Chemokines and their role in airway hyper-reactivity

Kate Blease, Nicholas W Lukacs, Cory M Hogaboam and Steven L Kunkel

University of Michigan Medical School, Ann Arbor, Michigan, USA

Received: 5 April 2000  
Revisions requested: 15 May 2000  
Revisions received: 20 June 2000  
Accepted: 20 June 2000  
Published: 5 July 2000

*Respir Res* 2000, 1:54-61

© Current Science Ltd (Print ISSN 1465-9921; Online ISSN 1465-993X)

## Abstract

Airway hyper-reactivity is a characteristic feature of many inflammatory lung diseases and is defined as an exaggerated degree of airway narrowing. Chemokines and their receptors are involved in several pathological processes that are believed to contribute to airway hyper-responsiveness, including recruitment and activation of inflammatory cells, collagen deposition and airway wall remodeling. These proteins are therefore thought to represent important therapeutic targets in the treatment of airway hyper-responsiveness. This review highlights the processes thought to be involved in airway hyper-responsiveness in allergic asthma, and the role of chemokines in these processes. Overall, the application of chemokines to the prevention or treatment of airway hyper-reactivity has tremendous potential.

**Keywords:** asthma, eosinophils, fibrosis, T cells

## Introduction

One of the key features of pulmonary diseases such as allergic asthma, cystic fibrosis and chronic obstructive pulmonary disease is the development of airway hyper-responsiveness (AHR) [1-3]. The factors involved in the development of AHR seem to differ between diseases, so for clarity this review will focus on the development of AHR during allergic asthmatic disease. In the context of asthma, AHR equates to an exaggerated bronchoconstrictor response, not only to allergens to which the subjects are sensitized, but also to a range of non-specific stimuli, including agents as diverse as cold air and methacholine.

Under normal conditions, airway reactivity, the ability to alter the size of the airways reversibly in response to stimuli, is an essential component of homeostasis. For example, when there is a need to move large volumes of air, such as with exercise, bronchial dilation occurs. Conversely, when it is important to limit or decrease the volume of air inspired, such as with exposure to irritating gases, the lung defends itself with coughing and bronchial narrowing. When this response is excessive, it is referred to as airway or bronchial hyper-reactivity or hyper-responsiveness (AHR) and manifests itself as an exaggerated bronchoconstrictor response to various provocative

AHR = airway hyper-responsiveness; BALF = bronchoalveolar lavage fluid; CCR = CC chemokine receptor; CXCR = CXC chemokine receptor; FEV<sub>1</sub> = forced expiratory volume over 1 s; IL = interleukin; LTC<sub>4</sub> = leukotriene C<sub>4</sub>; MCP = macrophage chemoattractant protein; MIP = macrophage inflammatory protein; PC<sub>20</sub> = provocative concentration; PD<sub>20</sub> = provocative dose; Th cells, T helper cells.

**Table 1**

**Chemokine receptors and their ligands**

CXC chemokine receptors	CC chemokine receptors
CXCR1: IL-8, GCP-2	CCR1: MIP-1 $\alpha$ , RANTES, MCP-3, MIP-5
CXCR2: IL-8, GCP-2, GRO $\alpha$ , $\beta$ , $\gamma$ , ENA-78	CCR2: MCP-1 to MCP-5
CXCR3: IP-10, MIG, ITAC	CCR3: Eotaxin, MCP-3,4, RANTES
CXCR4: SDF-1	CCR4: TARC, MDC
CXCR5: BCA	CCR5: MIP-1 $\alpha$ , RANTES, MIP-1 $\beta$
	CCR6: LARC, MIP-3 $\alpha$
	CCR7: SLC, MIP-3 $\beta$
	CCR8: I-309, TARC, MIP-1 $\beta$
	CCR9: MIP-1 $\alpha$ , $\beta$ , MCP-1, MCP-5
	CCR10: SLC, LARC, BLC-1, ESKine
	CCR11: MCP-1 to MCP-5, eotaxin

GCP-2, granulocyte chemotactic protein-2; GRO, growth-related oncogene; IP-10,  $\gamma$ -interferon-inducible protein 10; MIG, monokine-induced by  $\gamma$ -interferon; TARC = T cell and activation-related chemokine; SLC = secondary lymphoid tissue chemokine; SDF-1, stromal cell-derived factor; BCA, B-cell chemoattractant; ENA, epithelial cell-derived neutrophil-activating factor; RANTES, regulated upon activation normal T cell expressed and secreted; ITAC, interferon-inducible T cell  $\alpha$  chemoattractant.

agents. Measurements of AHR have traditionally been used to identify individuals who are at risk of developing asthma or related illnesses. The essential feature to these tests is to provide stimuli of varying intensity, such as methacholine, to the airways of the individual and record the decrease in lung function that develops. The resulting stimulus-response curve that develops is then analyzed to determine the quantity of agent required to produce a given degree of obstruction as measured by various spirometric or plethysmographic variables. Such changes are usually expressed as a percentage decrease in forced expiratory volume over 1 s (FEV<sub>1</sub>). The three variables that are most often examined in quantifying the magnitude of the response are the concentration of an agonist that induces a fixed decrease in lung function (ie a 20% decrease in FEV<sub>1</sub>), the slope of the dose-response curve, and the dose at which a plateau can be produced. Typically, the response is expressed as either a provocative dose (PD<sub>20</sub>) or a provocative concentration (PC<sub>20</sub>).

How this hyper-responsive state is acquired is poorly understood; however, in general, as the disease process becomes more severe the airways become more responsive. At present it is believed that AHR can result from the coordination of several mechanisms, some or all of which might be operative in individual asthmatics. In asthma a relationship seems to exist between the inflammatory state of the airways and the severity of hyper-responsiveness. In addition, airway remodeling, including smooth muscle hyperplasia/hypertrophy, collagen deposition and sub-epithelial fibrosis, might contribute to the development of AHR [4-6]. Because recent work in the field of chemokine

biology has highlighted a role for these proteins in many of these inflammatory processes, chemokines might be intimately involved in the initiation and maintenance of AHR. In this regard, chemokines could be attractive therapeutic targets for the treatment of pulmonary disease with an AHR component, in particular asthma.

### Introduction to chemokines

During the past decade, our understanding of the mechanisms involved in the initiation and maintenance of pulmonary disease has been greatly aided by advances in the field of chemokine biology. Chemokines comprise four supergene families, classified into groups on the basis of the number and arrangement of conserved amino acid sequences at the N-terminus. Two of these families (the CC and CXC chemokine groups) contain over 50 identified ligands and at least 14 individual receptors (Table 1). Two additional chemokine families (C and CX<sub>3</sub>C chemokines) are small and contain, respectively, lymphotactin and fractalkine as their members. Recent knowledge of this superfamily has grown significantly as a result of the availability of large databases of expressed sequence tags and bioinformatics [7]. Furthermore, characterization of these chemokines *in vivo* has identified multiple roles within inflammation, including the regulation of leukocyte trafficking, the immunomodulation of leukocyte activation, fibrosis, angiogenesis, hematopoiesis and organogenesis [8].

The biological effects of chemokines are mediated by the interaction of these soluble proteins with specific receptors, which belong to the superfamily of seven-transmembrane G-protein-coupled receptors. So far, 11 CC chemokine

receptors, five CXC chemokine receptors, one CX<sub>3</sub>C chemokine receptor and one C chemokine receptor have been characterized [7,9]. Chemokine receptors exhibit multiple ligand specificity, although the chemokine-ligand promiscuity does not usually cross the boundaries between CC and CXC, except for the promiscuous duffy antigen receptor complex that is believed to act as a sink for unbound chemokines. Chemokine receptor distribution on leukocytes confers selective chemoattractant activities for leukocyte subsets, making them ideal candidates for a role in leukocyte subset trafficking at sites of inflammation; that is, getting the correct subpopulation of cells to migrate into the tissue. Whereas CXC chemokines such as interleukin-8 (IL-8) activate predominantly neutrophils, CC chemokines such as RANTES and eotaxin target a variety of cell types including macrophages, eosinophils and basophils. However, controversial results have been published regarding this distinct chemokine receptor profile on leukocytes, particularly in allergic diseases. It has been shown recently that, after the appropriate stimuli, the CC chemokine receptors CCR1 and CCR3 can be expressed on neutrophils, indicating a wider role for CC chemokines than mononuclear cell activation and recruitment [10,11]. Furthermore, both the CXC chemokine receptors CXCR1 and CXCR2 have been identified on eosinophils in addition to neutrophils [12]. However, chemokine receptor expression is not limited to inflammatory cells. It is interesting to note that structural cells such as epithelial cells, endothelial cells, smooth muscle cells and fibroblasts also express chemokine receptors and are able to produce chemokines; they are therefore capable of contributing to a wide range of biological functions [13-15].

Once chemokines are released, they can have profound and longlasting biological effects both in the microenvironment of their release and at distant sites. These effects, including leukocyte recruitment and activation, smooth muscle proliferation, regulation of collagen deposition and coordination of fibrosis, might have key individual roles in the establishment and maintenance of AHR [4,5].

### Chemokines and leukocyte recruitment in AHR

Studies in both animals and humans have demonstrated a positive correlation between the inflammatory state of the airways and the severity of AHR. However, because the type and cause of this inflammation, as well as the extent and consequences of the inflammatory process, vary between different diseases exhibiting AHR (Table 2), the direct contribution of individual cell types or chemokines to AHR is not yet clearly understood. As discussed above, the distinct pattern of chemokine receptors on leukocytes means that chemokines can exert effects on particular leukocyte subsets. Therefore, the selective recruitment of leukocytes to sites of inflammation in these diseases is strongly influenced by the temporal pattern of chemokine expression.

Table 2

Cellular infiltrate in the airway wall in asthma and chronic obstructive pulmonary disease

Asthma	Chronic obstructive pulmonary disease
T lymphocytes, CD4	T lymphocytes, CD8
CD25	CD25, VLA-1
Eosinophilia	Mild eosinophilia
Activated eosinophils	Non-activated eosinophils
Mast cells	Mast cells
Neutrophils	Neutrophils
	Macrophages

### Eosinophil recruitment and AHR

Lung eosinophilia is a fundamental trait of allergic asthma, and infiltration of the airways by eosinophils seems to be central in the pathogenesis of this disease [16-18]. Eosinophils and their products have been identified in sputum, bronchoalveolar lavage fluid (BALF) and biopsy material of the airways of patients with asthma. Furthermore, the number of these cells and the amount of their products correlate with the severity of airway reactivity [16,17,19,20].

Eosinophils contribute to the development of AHR through the activation, degranulation and release of proteases and oxidative products stored in their granules. These proteins include major basic protein (MBP), eosinophil cationic protein (ECP), eosinophil-derived neurotoxin (EDN) and eosinophil peroxidase (EPO). In addition, eosinophils generate oxidative products and lipid mediators, including platelet-activating factor and leukotriene C<sub>4</sub> (LTC<sub>4</sub>). The generation of these cytotoxic products can cause extensive tissue damage and enhance the accumulation of inflammatory cells. Damage to airway epithelium appears to correlate with airway hyper-reactivity because the loss of epithelium leads to the exposure of 'irritant' receptors of nerves, which might increase the response of the airways to various stimuli.

Several chemokines, including macrophage chemoattractant protein (MCP)-3, macrophage inflammatory protein (MIP)-1 $\alpha$ , MCP-4, RANTES and eotaxin, elicit the migration of eosinophils [21-23] and can confer some degree of selectivity on eosinophil recruitment. Specifically, eotaxin, a potent activator of eosinophils and T helper 2 (Th2) lymphocytes, interacts with CCR3 expressed on eosinophils [24-28] to cause both degranulation and chemotaxis of eosinophils [29,30]. Elevated levels of eotaxin detected in the sputum of asthmatics has been shown to be correlated with increased eosinophil numbers and eosinophil cationic protein levels [31]. In several

murine models of asthma, a pronounced lung eosinophilia was associated with an increase in eotaxin expression; a neutralizing antibody against eotaxin significantly inhibited eosinophil infiltration after antigen challenge and decreased AHR in these animals [28]. Contrasting effects on eosinophil recruitment and AHR have been demonstrated in eotaxin gene-deficient mice, possibly owing to the presence of the other, recently identified, CCR3-specific ligands eotaxin-2 and eotaxin-3 [32,33]. In addition to eosinophil chemotaxis and activation, eotaxin, in combination with IL-5, has been shown to mobilize eosinophils from the bone marrow, thereby increasing circulating numbers of eosinophils within the blood [34]. However, eotaxin is not the only chemokine able to modulate eosinophil accumulation within the lung. Murine models of allergic inflammation have shown the movement of eosinophils during the early stages of asthma to be dependent on RANTES and MIP-1 $\alpha$ , whereas eotaxin has been shown to be necessary for eosinophil accumulation during chronic stages of the response [35,36]. Therefore, to target chemokines for therapeutic intervention effectively it is essential to understand the temporal pattern of chemokine release.

### Chemokine-induced recruitment of Th2 cells and AHR

In addition to eosinophils, T cells constitute a large proportion of the inflammatory cells within the lungs of asthmatics. Indeed, T-cell-mediated immune responses are believed to be important contributors to AHR in asthmatic patients through the release of chemokines and cytokines that enhance lung inflammation, favor the production of IgE, activate eosinophils and mast cells, and directly enhance AHR [37–39]. The observation that T cells have a role in AHR is supported by findings that the transfer of T cells from a hyper-responsive mouse strain into a hypo-reactive strain induces non-specific airway reactivity [40]. Furthermore, a characterization of lymphocyte populations in asthmatics and non-asthmatics has demonstrated differences in T cell subtypes in biopsy specimens and BALF from patients with asthma: in asthmatics, significantly higher numbers of Th2-type cells were seen than in control subjects, whereas there was no difference in the number of Th1-type cells [41]. Th2-type cells can be distinguished by the profile of cytokines that they produce, such as IL-4, IL-13 and IL-5, which favor the production of IgE and the growth and activation of eosinophils and mast cells, in addition to enhancing AHR *in vivo* [37–39].

Although lymphocytes have long been known to accumulate at sites of immune and inflammatory reactions, attractants that induce these responses have been identified only recently. RANTES, MIP-1 $\alpha$  and MIP-1 $\beta$  were the first chemokines for which lymphocyte-chemotactic activity was reported. The monocyte-chemotactic proteins (MCP-1, MCP-2, MCP-3 and MCP-4) are also potent attractants

of T lymphocytes. Gonzalo *et al* [28], using neutralizing antibodies directed against MCP-1 or MCP-5, significantly attenuated the recruitment of both eosinophils and T cells to the lung in a murine model of ovalbumin-induced airway inflammation, and drastically reduced AHR. In contrast, the neutralization of MIP-1 $\alpha$  caused only a slight reduction in eosinophilia and AHR, and had no effect on T cell accumulation [28]. In a separate study by Lukacs *et al* [42], neutralization of MIP-1 $\alpha$  or RANTES had no effect on AHR in a murine model of allergy, although eosinophilia was reduced significantly.

The expression of chemokine receptors on lymphocytes and their responsiveness to chemokines vary considerably between subsets. CCR5 is expressed preferentially in Th1 cells, whereas CCR3 and CCR4 seem to be characteristic of Th2 cells [43,44]. It is therefore not surprising that chemokines that preferentially recruit Th2-type cells have recently been identified. A number of chemokines have been shown to have the ability to recruit Th2-type cells preferentially, including monocyte-derived chemokine (MDC) and I-309 [45,46]. T cells recruited to the lung by these chemokines may regulate the persistence and activation of other cells such as eosinophils or mast cells in the airways of patients with asthma via both direct contact and through the release of other inflammatory mediators which contribute to enhanced AHR.

### Mast cells and AHR

Mast cells that are located in mucosal and peribronchovascular areas of the lung are known to be important in allergic reactions within the lung. These cells have the capacity to release a variety of mediators that can cause acute bronchospasm, activate and/or attract other inflammatory cells in the lung, and possibly increase AHR [47]. Indeed, there is a strong correlation between amounts of histamine in the airways of allergic asthmatics and sensitivity of the airways to methacholine [48,49].

MCP-1, a CC chemokine that binds CCR2, has been shown to induce AHR by the activation of mast cells in the lung. Activation of mast cells with MCP-1 causes the release of histamine, leukotrienes, platelet-activating factor and various proteases that either directly mediate changes in AHR or further enhance the recruitment of leukocytes to the lungs [36]. Increased levels of MCP-1, in murine models of allergic inflammation, have been shown to activate mast cells directly [36]. In addition, increased levels of MCP-1 have been detected in BALF and bronchial tissue from patients with atopic asthma in comparison with controls [50,51]. With the use of a murine model of cockroach antigen-induced allergic airway inflammation, it has been demonstrated that anti-MCP-1 antibodies inhibit AHR to methacholine and attenuate histamine release into the BALF; furthermore, in normal mice, instillation of MCP-1 induced prolonged airway hyper-reactivity and

histamine release. In addition, MCP-1 directly induced pulmonary mast cell degranulation *in vitro* [36]. In asthmatic patients, histamine and LTC<sub>4</sub> either directly induce AHR or facilitate the recruitment of leukocytes to the lungs to induce AHR indirectly [52,53]. Thus, the induction and evolution of allergic airway inflammation which is dependent on the temporal expression of multiple chemokines and their ligands have been shown to play a key role in the establishment of AHR.

### The role of airway remodeling and subepithelial fibrosis in AHR

Although several studies show a direct correlation between AHR and airway inflammation, the causal relationship between leukocyte infiltration and AHR has not been finally settled. There is a discordance in the findings between investigative groups who have studied the relationship between airway inflammation, as assessed by cellular infiltration, and AHR. Some groups have shown a strong relationship between the presence of inflammatory cells and enhanced airway responsiveness [16,17,19,20], whereas other groups have failed to establish such a relationship [18,54–57]. The conflicting evidence might reflect the reality that other factors in addition to, or distinct from, airway inflammation may modulate AHR. Of particular interest is the role that airway remodeling and subepithelial fibrosis play in AHR.

### Airway wall thickening, airway smooth muscle hypertrophy and subepithelial fibrosis in AHR

Histologic studies have reported a marked increase in the amount of smooth muscle in airways from asthmatic subjects; this abnormality, along with airway inflammation, is thought to contribute to AHR. It is believed that increased smooth muscle mass would allow the development of greater force and enhanced narrowing of the airway lumen to a given contractile stimulus. It has also been shown that smooth muscle cells can display different phenotypes depending on their environment or culture conditions. Smooth muscle cells have been shown to exhibit a classical contractile phenotype and also a proliferative-synthetic phenotype, which are capable of producing pro-inflammatory cytokines, chemokines and growth factors that further affect the environment within the lung [58]. Airway smooth muscle cells releasing chemokines such as eotaxin, RANTES, MCP-1, MCP-2 and MCP-3 [59–61] augment inflammatory responses within the lung such as leukocyte recruitment and activation, as discussed previously, that further exacerbate AHR. Because the increase in bronchial smooth muscle mass in asthma is due to cell hypertrophy in addition to hyperplasia [62], the potential relevance of phenotype plasticity and its possible relationship to altered function of smooth muscle in disease states has been suggested. Allergic sensitization and exposure to certain cytokines elicit significant functional changes [39,63,64] that can alter both contractile and

Table 3

Elevated chemokines in allergic asthmatic lungs shown *in vivo* to participate in AHR

Chemokine elevated in allergic asthmatics	<i>In vivo</i> evidence for involvement in AHR so far
Eotaxin	Yes
Eotaxin-2	No
RANTES	Yes
MCP-3	No
MCP-1	Yes
MIP-1 $\alpha$	Yes
CCR3	Yes

secretory functions; however, it remains to be seen how chemokines can alter this phenotype.

Although fibrosis is an essential component of tissue healing and wound repair, clinical studies have demonstrated that the degree of subepithelial fibrosis is correlated with augmented AHR to methacholine [4]. Indeed, a buildup of interstitial collagen beneath the airway basement membrane and subepithelial fibrosis are present in the airways of allergic asthmatics [6]. Infiltrating inflammatory cells such as macrophages, lymphocytes, neutrophils and eosinophils participate in the pathogenesis of lung fibrosis, through the activation of fibroblasts via the release of inflammatory mediators or direct contact [65,66]. Recent evidence has shown that MCP-1 enhances collagen deposition by fibroblasts [67]; therefore, increased expression of this chemokine in the lungs of asthmatics might be responsible for the airway remodeling that can exacerbate AHR.

### Chemokine expression in allergic asthma and their therapeutic use

So far, most of the results indicating a role for chemokines in AHR have been obtained through murine models employing chemokine neutralization, transgenic methods or gene knockout methods. The question therefore arises as to why chemokines would be beneficial targets for the therapeutic treatment of AHR in humans. Clinical studies have shown elevated levels of the chemokines and chemokine receptors that have been identified in murine models and in BALF, bronchial biopsies and sputum from allergic asthmatics (Table 3). Eotaxin, CCR3, mRNA and protein have been found to be significantly elevated in bronchial mucosal biopsies from atopic asthmatics in comparison with normal controls [68]. Furthermore, an inverse correlation was made in this study between the expression of eotaxin mRNA and the histamine provoca-

tive concentration causing a 20% decrease in FEV<sub>1</sub> [68]. Significant correlations with clinical parameters of AHR were also found with MCP-1 levels in BALF from patients with allergic asthma [69]. A comprehensive study by Ying *et al* [70] measured elevated levels of mRNA for eotaxin, eotaxin-2, RANTES, MCP-3, MCP-4 and CCR3 in the bronchial mucosa from allergic asthmatics [67]. In addition, levels of RANTES, MIP-1 $\alpha$  and MCP-1 in BALF have been shown to be significantly increased 4 h after challenge with endobronchial allergen in allergic asthmatics compared with levels before the allergen challenge [71]. Therefore the chemokines that have been shown to have a role in AHR in murine models are elevated in human disease and might be potential targets for the development of therapeutic interventions.

## Conclusions

Taken together, both experimental evidence from murine models and clinical evidence of elevated chemokine and chemokine receptor levels in the allergic asthmatic lung suggest that chemokines, and their receptors, seem to be effective targets for the development of therapeutic interventions to be used in addition to current therapy for the treatment of AHR. However, it remains to be seen whether the first clinical trials bear out this promise.

## References

Articles of particular interest have been highlighted as:

- of special interest
- of outstanding interest

1. McDowell KM: Pathophysiology of asthma. *Respir Care Clin N Am* 2000, 6:15-28.
2. Tepper RS, Eigen H: Airway reactivity in cystic fibrosis. *Clin Rev Allergy* 1991, 9:159-168.
3. Postma DS, Kerstjens HA: Characteristics of airway hyperresponsiveness in asthma and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 1998, 158:S187-S192.
4. Boulet LP, Laviolette M, Turcotte H, Cartier A, Dugas M, Malo JL, Boutet M: Bronchial subepithelial fibrosis correlates with airway responsiveness to methacholine. *Chest* 1997, 112:45-52.
5. Boulet LP, Chakir J, Dube J, Laprise C, Boutet M, Laviolette M: Airway inflammation and structural changes in airway hyperresponsiveness and asthma: an overview. *Can Respir J* 1998, 5:16-21.
6. Elias JA, Zhu Z, Chupp G, Homer RJ: Airway remodeling in asthma. *J Clin Invest* 1999, 104:1001-1008.
7. Zlotnik A, Morales J, Hadrick JA: Recent advances in chemokines and chemokine receptors. *Crit Rev Immunol* 1999, 19:1-47.
8. Taub DD, Oppenheim JJ: Chemokines, inflammation and the immune system. *Ther Immunol* 1994, 1:229-248.
9. Schweickart VL, Epp A, Raport CJ, Gray PW: CCR11 is a functional receptor for the monocyte chemoattractant protein family of chemokines. *J Biol Chem* 2000, 275:9550-9558.
10. Bonocchi R, Polentarutti N, Luini W, Boreatti A, Bernasconi S, Locati M, Power C, Proudfoot A, Wells TN, Mackay C, Mantovani A, Sozzani S: Up-regulation of CCR1 and CCR3 and induction of chemotaxis to CC chemokines by IFN-gamma in human neutrophils. *J Immunol* 1999, 162:474-479.
11. Zhang S, Youn BS, Gao JL, Murphy PM, Kwon BS: Differential effects of leukotactin-1 and macrophage inflammatory protein-1 alpha on neutrophils mediated by CCR1. *J Immunol* 1999, 162:4938-4942.
12. Petering H, Gotze O, Kimmig D, Smolarski R, Kapp A, Elsner J: The biologic role of Interleukin-8: functional analysis and expression of CXCR1 and CXCR2 on human eosinophils. *Blood* 1999, 93:694-702.
13. Strieter RM, Phan SH, Showell HJ, Remick DG, Lynch JP, Genord M, Raiford C, Eskandari M, Marks RM, Kunkel SL: Monokine-induced neutrophil chemotactic factor gene expression in human fibroblasts. *J Biol Chem* 1989, 264:10621-10626.
14. Strieter RM, Kunkel SL, Showell HJ, Remick DG, Phan SH, Ward PA, Marks RM: Endothelial cell gene expression of a neutrophil chemotactic factor by TNF-alpha, LPS, and IL-1 beta. *Science* 1989, 243:1487-1489.
15. Standiford TJ, Kunkel SL, Basha MA, Chensue SW, Lynch JP 3d, Toews GB, Westwick J, Strieter RM: Interleukin-8 gene expression by a pulmonary epithelial cell line. A model for cytokine networks in the lung. *J Clin Invest* 1990, 86:1945-1953.
16. Bentley AM, Manz G, Storz C, Robinson DS, Bradley B, Jeffery PK, Durham SR, Kay AB: Identification of T lymphocytes, macrophages, and activated eosinophils in the bronchial mucosa in intrinsic asthma. Relationship to symptoms and bronchial responsiveness. *Am Rev Respir Dis* 1992, 146:500-508.
17. Wardlaw AJ, Dunnette S, Gleich GJ, Collins JV, Kay AB: Eosinophils and mast cells in bronchoalveolar lavage in subjects with mild asthma. Relationship to bronchial hyperreactivity. *Am Rev Respir Dis* 1988, 137:82-89.
18. Ferguson AC, Wong FW: Bronchial hyperresponsiveness in asthmatic children. Correlation with macrophages and eosinophils in bronchoalveolar fluid. *Chest* 1989, 96:988-991.
19. Azzawi M, Bradley B, Jeffery PK, Frew AJ, Wardlaw AJ, Knowles G, Asoufi B, Collins JV, Durham S, Kay AB: Identification of activated T lymphocytes and eosinophils in bronchial biopsies in stable atopic asthma. *Am Rev Respir Dis* 1990, 142:1407-1413.
20. Kirby JG, Hargreave FE, Gleich GJ, O'Byrne PM: Bronchoalveolar cell profiles of asthmatic and nonasthmatic subjects. *Am Rev Respir Dis* 1987, 136:379-383.
21. Lukacs NW, Standiford TJ, Chensue SW, Kunkel RG, Strieter RM, Kunkel SL: C-C chemokine-induced eosinophil chemotaxis during allergic airway inflammation. *J Leukoc Biol* 1998, 60:573-578.
22. Griffiths-Johnson DA, Collins PD, Rossi AG, Jose PJ, Williams TJ: The chemokine, eotaxin, activates guinea-pig eosinophils *in vitro* and causes their accumulation into the lung *in vivo*. *Biochem Biophys Res Commun* 1993, 197:1167-1172.
23. Stafford S, Li H, Forsythe PA, Ryan M, Bravo R, Alam R: Monocyte chemoattractant protein-3 (MCP-3)/fibroblast-induced cytokine (FIC) in eosinophilic inflammation of the airways and the inhibitory effects of an anti-MCP-3/FIC antibody. *J Immunol* 1997, 158:4953-4960.
24. Teixeira MM, Wells TN, Lukacs NW, Proudfoot AE, Kunkel SL, Williams TJ, Hellewell PG: Chemokine-induced eosinophil recruitment. Evidence of a role for endogenous eotaxin in an *in vivo* allergy model in mouse skin. *J Clin Invest* 1997, 100:1657-1666.
25. Rothenberg ME, Ownbey R, Mehlhop PD, Loiselle PM, van de Rijn M, Bonventre JV, Oettgen HC, Leder P, Luster AD: Eotaxin triggers eosinophil-selective chemotaxis and calcium flux via a distinct receptor and induces pulmonary eosinophilia in the presence of interleukin 5 in mice. *Mol Med* 1996, 2:334-348.
26. Daugherty BL, Siciliano SJ, DeMartino JA, Malkowitz L, Sirotna A, Springer MS: Cloning, expression, and characterization of the human eosinophil eotaxin receptor. *J Exp Med* 1996, 183:2349-2354.

27. Ponath PD, Qin S, Ringler DJ, Clark-Lewis I, Wang J, Kassam N, Smith H, Shi X, Gonzalo JA, Newman W, Gutierrez-Ramos JC, Mackay CR: Cloning of the human eosinophil chemoattractant, eotaxin. Expression, receptor binding, and functional properties suggest a mechanism for the selective recruitment of eosinophils. *J Clin Invest* 1996, 97:604-612.
28. Gonzalo JA, Lloyd CM, Kremer L, Finger E, Martinez-A C, Siegelman MH, Cybulsky M, Gutierrez-Ramos JC: Eosinophil recruitment to the lung in a murine model of allergic inflammation. The role of T cells, chemokines, and adhesion receptors. *J Clin Invest* 1996, 98:2332-2345.
29. Kampen GT, Stafford S, Adachi T, Jinquan T, Quan S, Grant JA, Skov PS, Poulsen LK, Alam R: Eotaxin induces degranulation and chemotaxis of eosinophils through the activation of ERK2 and p38 mitogen-activated protein kinases. *Blood* 2000, 95:1911-1917.
30. Honda K, Chihara J: Eosinophil activation by eotaxin - eotaxin primes the production of reactive oxygen species from eosinophils. *Allergy* 1999, 54:1262-1269.
31. Yamada H, Yamaguchi M, Yamamoto K, Nakajima T, Hirai K, Morita Y, Sano Y, Yamada H: Eotaxin in induced sputum of asthmatics: relationship with eosinophils and eosinophil cationic protein in sputum. *Allergy* 2000, 55:392-397.
32. Forssmann U, Ugucioni M, Loetscher P, Dahinden CA, Langen H, Thelen M, Baggiolini M: Eotaxin-2, a novel CC chemokine that is selective for the chemokine receptor CCR3, and acts like eotaxin on human eosinophil and basophil leukocytes. *J Exp Med* 1997, 185:2171-2176.
33. Shinkai A, Yoshisue H, Koike M, Shoji E, Nakagawa S, Saito A, Takeda T, Imabeppu S, Kato Y, Hanai N, Anazawa H, Kuga T, Nishi T: A novel human CC chemokine, eotaxin-3, which is expressed in IL-4-stimulated vascular endothelial cells, exhibits potent activity toward eosinophils. *J Immunol* 1999, 163:1602-1610.
34. Palframan RT, Collins PD, Williams TJ, Rankin SM: Eotaxin induces a rapid release of eosinophils and their progenitors from the bone marrow. *Blood* 1998, 91:2240-2248.
35. Campbell EM, Kunkel SL, Strieter RM, Lukacs NW: Temporal role of chemokines in a murine model of cockroach-allergen-induced airway hyperreactivity and eosinophilia. *J Immunol* 1998, 161:7047-7053.
36. Campbell EM, Charo IF, Kunkel SL, Strieter RM, Boring L, Gosling J, Lukacs NW: Monocyte chemoattractant protein-1 mediates cockroach allergen-induced bronchial hyperreactivity in normal but not CCR2-/- mice: the role of mast cells. *J Immunol* 1999, 163:2180-2187.
37. Bradley BL, Azzawi M, Jacobson M, Assoufi B, Collins JV, Irani AM, Schwartz LB, Durham SR, Jeffery PK, Kay AB: Eosinophils, T-lymphocytes, mast cells, neutrophils, and macrophages in bronchial biopsy specimens from atopic subjects with asthma: comparison with biopsy specimens from atopic subjects without asthma and normal control subjects and relationship to bronchial hyperresponsiveness. *J Allergy Clin Immunol* 1991, 88:661-674.
38. Shi HZ, Deng JM, Xu H, Nong ZX, Xiao CQ, Liu ZM, Qin SM, Jiang HX, Liu GN, Chen YC: Effect of inhaled interleukin-4 on airway hyperreactivity in asthmatics. *Am J Respir Crit Care Med* 1998, 157:1818-1821.
39. Hakonarson H, Maskeri N, Carter C, Chuang S, Grunstein MM: Autocrine interaction between IL-5 and IL-13 mediates altered responsiveness of atopic asthmatic sensitized airway smooth muscle. *J Clin Invest* 1999, 104:657-667.
40. De Sanctis GT, Itoh A, Green FH, Qin S, Kimura T, Grobholz JK, Martin TR, Maki T, Drazen JM: T-lymphocytes regulate genetically determined airway hyperresponsiveness in mice. *Nat Med* 1997, 3:460-462.
41. Robinson DS, Hamid Q, Ying S, Tsicopoulos A, Barkans J, Bentley AM, Corrigan C, Durham SR, Kay AB: Predominant TH2-like bronchoalveolar T-lymphocyte population in atopic asthma. *N Engl J Med* 1992, 326:298-304.
42. Lukacs NW, Strieter RM, Warrington K, Lincoln P, Chensue SW, Kunkel SL: Differential recruitment of leukocyte populations and alteration of airway hyperreactivity by C-C family chemokines in allergic airway inflammation. *J Immunol* 1997, 158:4398-4404. These authors demonstrate the involvement of multiple CC chemokines in the recruitment of particular leukocyte populations to the lung during allergic airway inflammation and show evidence that neutralization of MCP-1 decreases AHR.
43. Bonocchi R, Bianchi G, Bordinon PP, D'Ambrosio D, Lang R, Borzatti A, Sozzani S, Allavena P, Gray PA, Mantovani A, Sinigaglia F: Differential expression of chemokine receptors and chemotactic responsiveness of type 1 T helper cells (Th1s) and Th2s. *J Exp Med* 1998, 187:129-134.
44. Sallusto F, Mackay CR, Lanzavecchia A: Selective expression of the eotaxin receptor CCR3 by human T helper 2 cells. *Science* 1997, 277:2005-2007.
45. Lloyd CM, Delaney T, Nguyen T, Tian J, Martinez-A C, Coyle AJ, Gutierrez-Ramos JC: CC chemokine receptor (CCR)3/eotaxin is followed by CCR4/monocyte-derived chemokine in mediating pulmonary T helper lymphocyte type 2 recruitment after serial antigen challenge in vivo. *J Exp Med* 2000, 191:265-274.
46. Zingoni A, Soto H, Hedrick JA, Stoppacciaro A, Storazzi CT, Sinigaglia F, D'Ambrosio D, O'Garra A, Robinson D, Rocchi M, Santoni A, Zlotnik A, Napolitano M: The chemokine receptor CCR8 is preferentially expressed in Th2 but not Th1 cells. *J Immunol* 1998, 161:547-551.
47. Schulman ES: The role of mast cell derived mediators in airway hyperresponsiveness. *Chest* 1986, 90:578-583.
48. Casale TB, Wood D, Richerson HB, Zehr B, Zavala D, Hunninghake GW: Direct evidence of a role for mast cells in the pathogenesis of antigen-induced bronchoconstriction. *J Clin Invest* 1987, 80:1507-1511.
49. Casale TB, Wood D, Richerson HB, Trapp S, Metzger WJ, Zavala D, Hunninghake GW: Elevated bronchoalveolar lavage fluid histamine levels in allergic asthmatics are associated with methacholine bronchial hyperresponsiveness. *J Clin Invest* 1987, 79:1197-1203.
50. Sousa AR, Lane SJ, Nakhosteen JA, Yoshimura T, Lee TH, Poston RN: Increased expression of the monocyte chemoattractant protein-1 in bronchial tissue from asthmatic subjects. *Am J Respir Cell Mol Biol* 1994, 10:142-147.
51. Jahnz-Rozky KM, Kuna P, Pirozynska E: Monocyte chemotactic and activating factor/monocyte chemoattractant protein (MCAF/MCP-1) in bronchoalveolar lavage fluid from patients with atopic asthma and chronic bronchitis. *J Invest Allergol Clin Immunol* 1997, 7:254-259.
52. de Gouw HW, Verbruggen MB, Twiss IM, Sterk PJ: Effect of oral L-arginine on airway hyperresponsiveness to histamine in asthma. *Thorax* 1999, 54:1033-1035.
53. Crowther SD, Morley J, Costello JF: Varied effects of regular salbutamol on airway responsiveness to inhaled spasmogens. *Lancet* 1997, 350:1450.
54. Adenot E, Rosenhall L, Johansson SA, Linden M, Venge P: Inflammatory cells and eosinophilic activity in asthmatics investigated by bronchoalveolar lavage. The effects of antiasthmatic treatment with budesonide or terbutaline. *Am Rev Respir Dis* 1990, 142:91-99.
55. Ollerenshaw SL, Woolcock AJ: Characteristics of the inflammation in biopsies from large airways of subjects with asthma and subjects with chronic airflow limitation. *Am Rev Respir Dis* 1992, 145:922-927.



56. Djukanovic R, Wilson JW, Britten KM, Wilson SJ, Walls AF, Roche WR, Howarth PH, Holgate ST: Quantitation of mast cells and eosinophils in the bronchial mucosa of symptomatic atopic asthmatics and healthy control subjects using immunohistochemistry. *Am Rev Respir Dis* 1990, 142:863-871.
57. Jeffery PK, Wardlaw AJ, Nelson FC, Collins JV, Kay AB: Bronchial biopsies in asthma. An ultrastructural, quantitative study and correlation with hyperreactivity. *Am Rev Respir Dis* 1989, 140:1745-1753.
58. Johnson SR, Knox AJ: Synthetic functions of airway smooth muscle in asthma. *Trends Pharmacol Sci* 1997, 18:288-292.
59. John M, Hirst SJ, Jose PJ, Robichaud A, Berkman N, Witt C, Twort CH, Barnes PJ, Chung KF: Human airway smooth muscle cells express and release RANTES in response to T helper 1 cytokines: regulation by T helper 2 cytokines and corticosteroids. *J Immunol* 1997, 158:1841-1847.
60. Chung KF, Patel HJ, Fadlon EJ, Rousell J, Haddad EB, Jose PJ, Mitchell J, Belvisi M: Induction of eotaxin expression and release from human airway smooth muscle cells by IL-1 beta and TNF alpha: effects of IL-10 and corticosteroids. *Br J Pharmacol* 1999, 127:1145-1150.
61. Pype JL, Dupont LJ, Menten P, Van Coillie E, Opdenakker G, Van Damme J, Chung KF, Demedts MG, Verleden GM: Expression of monocyte chemoattractant protein (MCP)-1, MCP-2, and MCP-3 by human airway smooth-muscle cells. Modulation by corticosteroids and T-helper 2 cytokines. *Am J Respir Cell Mol Biol* 1999, 21:528-538.
62. Ebina M, Takahashi T, Chiba T, Motomiya M: Cellular hypertrophy and hyperplasia of airway smooth muscles underlying bronchial asthma. A 3-D morphometric study. *Am Rev Respir Dis* 1993, 148:720-726.
63. Mitchell RW, Ruhlmann E, Magnussen H, Löff AR, Rabe KF: Passive sensitization of human bronchi augments smooth muscle shortening velocity and capacity. *Am J Physiol* 1994, 267:L218-L222.
64. Mitchell RW, Rabe KF, Magnussen H, Löff AR: Passive sensitization of human airways induces myogenic contractile responses *in vitro*. *J Appl Physiol* 1997, 83:1276-1281.
65. Hogaboam CM, Smith RE, Kunkel SL: Dynamic interactions between lung fibroblasts and leukocytes: implications for fibrotic lung disease. *Proc Assoc Am Physicians* 1998, 110:313-320.
66. Halene M, Lake-Bullock V, Zhu J, Hao H, Cohen DA, Kaplan AM: T cell-independence of bleomycin-induced pulmonary fibrosis. *J Leukoc Biol* 1999, 65:187-195.
67. Gharaee-Kermani M, Denholm EM, Phan SH: Costimulation of fibroblast collagen and transforming growth factor beta 1 gene expression by monocytes chemoattractant protein-1 via specific receptors. *J Biol Chem* 1998, 271:17779-17784.
68. Ying S, Robinson DS, Meng Q, Rottman J, Kennedy R, Ringler DJ, Mackay CR, Daugherty BL, Springer MS, Durham SR, Williams TJ, Kay AB: Enhanced expression of eotaxin and CCR3 mRNA and protein in atopic asthma. Association with airway hyperresponsiveness and predominant co-localization of eotaxin mRNA to bronchial epithelial and endothelial cells. *Eur J Immunol* 1997, 27:3507-3516.
69. Rozyk KJ, Plusa T, Kuna P, Pirozynska E: Monocyte chemoattractant and activating factor/monocyte chemoattractant protein in bronchoalveolar lavage fluid from patients with atopic asthma and chronic bronchitis. Relationship to lung function tests, bronchial hyper-responsiveness and cytology of bronchoalveolar lavage fluid. *Immunol Lett* 1997, 58:47-52.
70. Ying S, Meng Q, Zeibecoglou K, Robinson DS, Macfarlane A, Humbert M, Kay AB: Eosinophil chemotactic chemokines (eotaxin, eotaxin-2, RANTES, monocyte chemoattractant protein-3 (MCP-3),

and MCP-4), and CC-chemokine receptor 3 expression in bronchial biopsies from atopic and nonatopic (intrinsic) asthmatics. *J Immunol* 1999, 163:6321-6328.

This study provides evidence that a variety of chemokines and their receptors are increased in bronchial biopsies from allergic asthmatic, correlating to results found in animal studies.

71. Holgate ST, Bodey KS, Janezic A, Frew AJ, Kaplan AP, Teran LM: Release of RANTES, MIP-1 alpha and MCP-1 into asthmatic airways following endobronchial allergen challenge. *Am J Respir Crit Care Med* 1997, 156:1377-1383.

**Authors' affiliation:** Department of Pathology, University of Michigan Medical School, Ann Arbor, Michigan, USA

**Correspondence:** Kate Bleese, PhD, Department of Pathology, University of Michigan Medical School, 5214 Med Sci I, 1301 Catherine Road, Ann Arbor, Michigan 48109, USA. Tel: +1 734 936 1020; fax: +1 734 764 2397; e-mail: kbleese@path.med.umich.edu

# The challenges of genome sequence annotation or "The devil is in the details"

Temple F. Smith and Xiaolin Zhang

Two powerful, competing pressures are acting on various genome sequencing projects: One, to release new sequences as quickly as possible; and two, to provide them with maximally complete and accurate annotation. This rather incongruent combination has led to a strong interest in developing efficient and accurate automated, large-scale sequence annotation procedures.

There have, in fact, been a number of attempts in both industry and academia to speed new sequence annotation. In their simplest form, these have been little more than post-processors acting on standard high-speed sequence similarity search tools such as BLAST. The post-processing assigns the annotation from the best-matched previously known sequence to each new sequence.

This is, of course, a generalization of successful approaches used by many researchers to assign probable functions to new sequences when previously studied and recognizable homologs exist. However, when applied in an automated manner to large data sets with minimum review, such approaches can lead to serious degradation of the wealth of incoming genomic data.

There are more problems with the simple best match functional annotation inheritance (BMAI) than the two traditionally recognized, those being the assessing of biological significance in terms of match statistical significance, and the choice between the sensitivity of the very fast, but approximate, sequence similarity search algorithms and the mathematically rigorous, but much slower, optimal algorithms.

In the first place, it is easy to assign various measures of confidence to new annotation based on match statistics, and there is good evidence that approximate maximum similarity tools such as BLAST do nearly as well as any of the slower, full dynamic programming methods. Second, the newer versions of BLAST have high sensitivity, identifying local sequence pairwise similarities, including alignment gaps. The inclusion of alignment gaps was one of the main advantages of the slower dynamic programming methods.

No, the major problems associated with nearly all of the current automated annotation approaches are—paradoxically—minor database annotation inconsistencies (and a few outright errors). This is particularly true for the large and often complex protein families. Why are these the major problems, rather than the two more obvious ones previously mentioned?

Clearly, for researchers studying a particular protein family, most database annotation inconsistencies make little difference in the search for new, even distant members. A local expert either knows the range and/or history of the annotation terminology used by colleagues in different subfields, or perhaps more importantly, the expert will spend the time to backtrack apparent inconsistencies.

Even in those cases involving structurally complex proteins composed of multiple domains, all of which may not be fully or properly annotated, the expert generally carefully dissects matches to distinct domains, and backtracks each domain's annotations. However, in the large-scale genomic projects, having a local expert to work on each protein family is not an option. Yet the integration of genomic information across multiple protein families, multiple fields of expertise and taxa, is just what is envisioned to form the foundations of the next century's biology and biotechnology.

The basic problem of inconsistent nomenclature arises largely because sequence information and its annotation derives from many diverse subdivisions of the biological sciences during a time of rapid change in our understanding. In an emerging field such as molecular biology, let alone "comparative genomics," strictly controlled vocabularies would not only be difficult to impose, but are probably undesirable! The evolution and refinement of the vocabulary is an anticipated outcome of our increasing knowledge.

Some inconsistencies are simple, such as the reference to tRNA synthetases in fungi as tRNA ligases (which of course they are) or the use by Americans and most Europeans of dihydroxyacetone-P for a glycolytic intermediate that the Japanese and English generally call glyceralone-P. There are many cases of equivalent, but different, terminology. For example, in the well-studied G protein case, among 27 distinct G  $\beta$ -subunit GenBank/SWISS-PROT entries, there are 18 different protein names or keyword sets. A list of synonyms can be constructed in such cases, some of which will be species or field specific.

There are numerous cases in which proteins of very different current functions are homologous in that they evolved from a common ancestor and will match with significant sequence similarity. For example, numerous proteins sharing multiple WD-repeats have been labeled transducin-like or transducin homologs, yet share no common signal transduction function! The rather widespread improper use of synthetase for synthase and the converse, however, cannot be fixed by a thesaurus, since whether the enzyme in question requires ATP or not is not a matter of alternate terminology. Without the careful use of synonym tables in

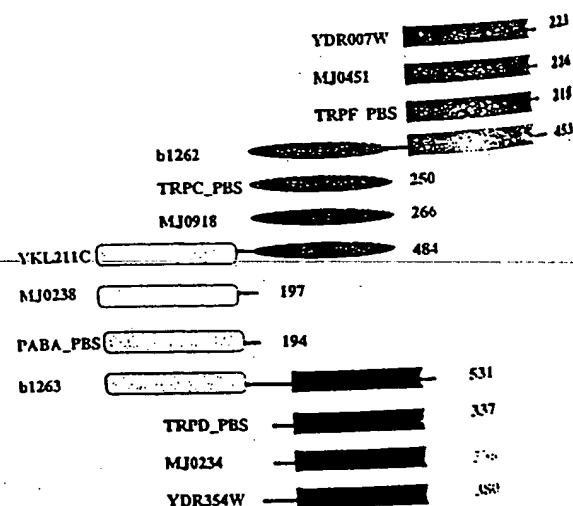


Figure 1. An example of genes having the potential for annotation inheritance transitivity. The three two-domain proteins, b1262, YKL211C, and b1263, share no single domain in common. Domains are labeled by colors: red, indole-3-phosphoribosyl anthranilate isomerase; green, glycerol phosphate synthase; yellow, anthranilate synthase subunit II; blue, anthranilate phosphoribosyltransferase.

Temple F. Smith is director and Xiaolin Zhang is a research associate at the BioMolecular Engineering Research Center, Boston University College of Engineering, 36 Cummington Street, Boston, MA 02215 (tsmith@darwin.bu.edu).

combination with review of commonly misused terminology, any simple BMAI approach will often end up propagating the less desirable or erroneous annotations.

Random propagation of faulty annotation, however, is only the tip of the annotation problem iceberg. In the case of multidomain proteins, most simple BMAI approaches will at best annotate only the most similar of the domains, and at worst will attach the annotation of a nonshared domain from the matched protein.

The first of these, incomplete annotation, is seen in the recently released *Escherichia coli* genome data for ORF b1262, a 453-residue, multifunctional protein<sup>1</sup>. Here, the first 253 amino acid residues comprise the indole-3-glycerol phosphate synthase domain. This matches single-domain homologs in *Methanococcus jannaschii* and *Bacillus subtilis* and the carboxy-terminal domain of the protein product of one yeast gene, YKL211C. The second domain of the *E. coli* protein residues 259 through 443 matches the *N*-phosphoribosyl anthranilate isomerase, single-domain protein in *M. jannaschii*, *B. subtilis*, and yeast (and this function is currently unannotated).

An incorrect inheritance via a matched multidomain protein is seen in the *M. jannaschii* ORF pair, MJ0234 and MJ0238. Both

match the *E. coli* ORF b1263, a bifunctional enzyme of two separate domains. Both *M. jannaschii* genes have been annotated, however, by only one of the two functions: anthranilate synthase subunit II, which is

**What must be done to avoid continued annotation inconsistency, incompleteness, and erroneous propagation?**

associated only with the first 176 of b1263's 531 amino acids, and that region is matched only by MJ0238 (Fig. 1).

What must be done to avoid continued annotation inconsistency, incompleteness, and erroneous propagation? First, any automation must be rather sophisticated. It must, for a start, recognize large differences in the length of matching sequences; it must associate annotation with specific subsequences; it must recognize all differences among the annotations of the homologs to the matched sequence; and, whenever possible, sequence similarity should be identified via shared conserved sequence patterns or profiles that have been

carefully annotated, consistent with the entire family characterized by that pattern. All approaches should exploit the best available synonym tables, such as those available through resources like PROSITE, the Enzyme Commission, or the US National Library of Medicine's UNLS database. Finally, any annotation strategy must be designed to support an evolving nomenclature and rapidly expanding knowledge base.

Even if it takes an extended period of time to annotate the new genome data more carefully and completely now, it will surely be more cost effective than redoing it later. Recall that the correcting and/or updating of all of the historical data in largely archival sequence databases such as GenBank or SWISS-PROT, has not yet been completed—probably for good reasons of cost and time. We in the basic research and biotechnology communities must not let our excitement or our impatience for the new data degrade its annotation and longer-term utility.

1. Neer, E.J., Schmidt, C.J., Nambudripad, R., and Smith, T.F. 1994. The ancient regulatory-protein family of WD-repeat proteins. *Nature* 371:297-300.
2. ORF data obtained from:  
*M. jannaschii*: [www.tigr.org/tdb/mdb/mdb.html](http://www.tigr.org/tdb/mdb/mdb.html)  
*E. coli*: [www.genetics.wisc.edu](http://www.genetics.wisc.edu)  
*S. cerevisiae*: [speedy.mips.biochem.mpg.de](http://speedy.mips.biochem.mpg.de)  
*B. subtilis*: [www.pasteur.fr/subti/subti\\_flat.html](http://www.pasteur.fr/subti/subti_flat.html)



# Shearwater Polymers, Inc.

## POLY(ETHYLENE GLYCOL) AND DERIVATIVES

### FUNCTIONALIZED BIOCOMPATIBLE POLYMERS FOR PROTEIN PEGYLATION, PHARMACEUTICAL AND BIOTECHNICAL APPLICATIONS

#### Amine Reactive Derivatives

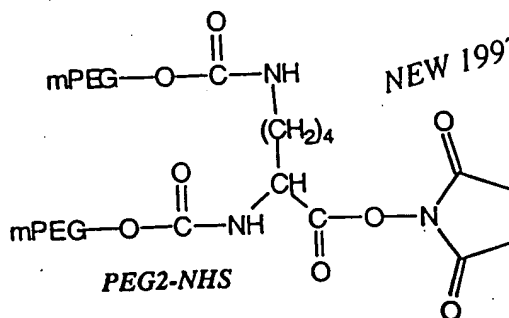
- Succinimidyl Succinate (SS)
- Succinimidyl Propionate (SPA)
- Succinimidyl Carboxymethyl (SCM)
- PEG2 Succinimide (PEG2-NHS)
- Oxycarbonylimidazole (CDI)
- Nitrophenyl carbonate (NPC)
- Tresylate (TRES)
- Epoxide (EPOX)
- Aldehyde (ALD)
- Isocyanate (NCO)

#### Heterofunctional Derivatives

- $\omega$ -amino- $\alpha$ -carboxyl (NH<sub>2</sub>COOH)
- $\omega$ -hydroxyl- $\alpha$ -amine (HONH<sub>2</sub>)
- $\omega$ -hydroxyl- $\alpha$ -carboxyl (HOCO<sub>2</sub>H)
- NHS-Vinylsulfone (NHSVS)
- NHS-Maleimide (NHSMAL)

#### Sulfhydryl Reactive Derivatives

- Vinyl Sulfone (VS)
- Maleimide (MAL)
- Orthopyridyl-disulfide (OPSS)



Call or write for full product listing  
(800) 457-1806 (U.S.)  
(205) 533-4201 (Int.)  
(205) 533-4805 (fax)  
2307 Spring Branch Rd.  
Huntsville, AL 35801  
[swpolymers@aol.com](mailto:swpolymers@aol.com) (e-mail)

Visit our web site at <http://www.swpolymers.com/>

#### NEW SERVICES AVAILABLE

PEGylation and Methods Development for your proteins, peptides, small-molecule drugs and other bioactive molecules

# Errors in genome annotation

At the time that Watson and Crick proposed a structure for DNA, a visionary might have suggested that the complete genetic sequence of an organism would eventually be known. However, nobody could have realistically proposed that machines could automatically indicate gene functions. Yet precisely this has been achieved: with no laboratory experiments at all, the roles of most genes in several organisms have been reported.

But how reliable are these functional assignments, upon which we depend for understanding genes and genomes? Without laboratory experiments to verify the computational methods and their expert analysis, it is impossible to know for certain. However, a simple procedure can place a rough upper bound on their accuracy. I have compared three different groups' functional annotation<sup>1-3</sup> for the *Mycoplasma genitalium* genome<sup>1</sup> (Fig. 1). Where two groups' descriptions are completely incompatible, at least one must be in error. In my analysis, there is no penalty

for vague or absent functional assignment. Furthermore, I always assume that as many groups as possible have the right description (Fig. 2).

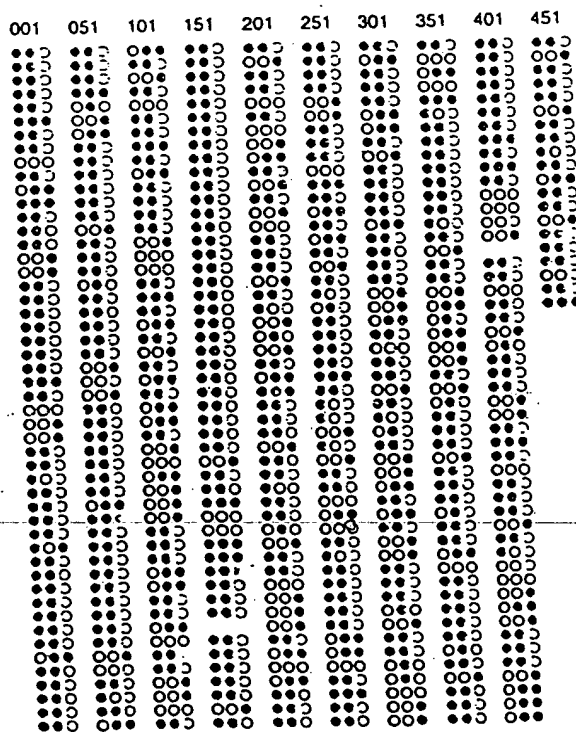
The results are disappointing for those expecting reliable annotation (Table 1). *M. genitalium* was reported to have just 468 genes, many of which are fundamental for all life and therefore easy to analyse. Nonetheless, the error rate is at least 8% for the 340 genes annotated by two or three groups. This value may not be uniform across the three groups: does it reflect the overall significance of a group's annotations? Genes annotated by only one group were not counted, but include such improbable bacterial functions as a sex enhancing factor, mitochondrial polymerase, and a coreceptor. This analysis cannot detect those cases where multiple groups arrived at consistent but wrong conclusions – a likely occurrence because all relied on the same methods and data. This evaluation also ignores minor disagreements in annotation, and disparities in description specificity (possibly indicating problematic over- or under-assignment of function<sup>4</sup>). Therefore, the true error rate may be greater than these figures indicate.

There are several possible reasons why the functional analyses have mistakes, as described at greater length elsewhere<sup>5-8</sup>. For example, it may be that the similarity between the genomic query and database sequence is insufficient to reliably detect homology, an issue solved by appropriate use of modern and accurate sequence comparison procedures<sup>9,10</sup>. A more difficult problem is the inference of function from homology. Typical database searching methods are valuable for finding evolutionarily related proteins, but if there are only about 1000 superfamilies in nature<sup>11,12</sup>, then most homologs will have different molecular and cellular functions.

The annotation problem escalates dramatically as the single genome, for genes with incorrect function entered into public databases. Subsequent searches against these databases then cause errors to propagate into future functional assignments. The procedure needs only a few times without corrections before the results that made computational function determination possible – the annotation databases – are so polluted as to be almost useless. To prevent errors from spreading, control, database curation by the scientific community will be essential<sup>6,13</sup>.

To ensure that databases are kept usable, the information for gene annotation should be clear: does it indicate a true ortholog, and/or functional equivalence? Fortunately, databases already incorporate this information (e.g. Ref. 14). Errors will, of course, still creep in, but eliminate the collateral damage, computational errors should clearly be flagged as such, and they should also indicate their source (which would allow for correction) and a measure of confidence in the annotation. This will require new research and development of algorithms and databases, and a broad commitment to maintaining these resources. In short, the access and documentation needed for reproducibility of a computational function determination should be commensurate with the effort for a corresponding laboratory bench experiment.

**FIGURE 1. Comparison of annotations**



Three dots represent (left to right) Frasier et al.<sup>1</sup>, Koonin et al.<sup>2</sup> and Ouzounis et al.<sup>3</sup> annotations for each of the 468 *M. genitalium* genes. (Tentative cases from Ouzounis et al.<sup>3</sup> were not used.) An open black circle indicates lack of a substantial functional annotation. Compatible annotations are colored identically, while conflicting annotations are in different colors. It is unknown which, if any, of the annotations are actually correct. There are 300 cases where Ouzounis et al.<sup>3</sup> simply reported the SWISS-PROT annotation of the same *M. genitalium* gene, indicated by colored open circles. Because Frasier et al.<sup>1</sup> annotation played a role in SWISS-PROT descriptions, these Ouzounis et al.<sup>3</sup> annotations were not included in this analysis. Though not incorporated in Table 1, the color indicates the compatibility of the functional annotation. The conflict/compatibility analysis here is itself certain to have errors; however, these should not affect the magnitude of the measured annotation error rate.

Steven E. Brenner  
brenner@hyper.  
stanford.edu

Department of Structural  
Biology, Stanford  
University, Fairchild  
Building, Stanford,  
CA 94305-5126, USA

**FIGURE 2. Example annotations and analysis**

<b>(a)</b>		<b>(b)</b>	
mg463		mg302	
Frasier <i>et al.</i>	• High level kasamycin resistance (ksgA)	Frasier <i>et al.</i>	○ No database match
Koonin <i>et al.</i>	• rRNA (adenosine-N6,N6-dimethyltransferase (ksoA))	Koonin <i>et al.</i>	• (Glycerol-3-phosphate?) permease
mg010		mg442	
Frasier <i>et al.</i>	• DNA primase (dnaE)	Frasier <i>et al.</i>	• Pili repressor (pilB)
Koonin <i>et al.</i>	• DNA primase (truncated version) (DnaGp)	Koonin <i>et al.</i>	• Putative chaperone like protein
Ouzounis <i>et al.</i>	• DNA primase (EC 2.7.7.-)	Ouzounis <i>et al.</i>	• PilB protein
mg225		mg085	
Frasier <i>et al.</i>	○ Hypothetical protein	Frasier <i>et al.</i>	• Hydroxymethylglutaryl-CoA reductase (NADPH)
Koonin <i>et al.</i>	• Amino acid permease	Koonin <i>et al.</i>	• ATP(GTP?)-utilizing enzyme
Ouzounis <i>et al.</i>	• Histidine permease	Ouzounis <i>et al.</i>	• NADH-ubiquinone oxidoredu (sic)

(a) Consistent annotations. Annotations were generally considered consistent for this analysis if either the function or the gene name match (e.g. mg463; mg010). An exception is when one group uses a gene name and another specifically notes that the current gene is a paralog and not identical (consider mg010). Where the descriptions from different groups were compatible, but of different levels of specificity, this was considered a correct assignment (e.g. mg225). The difficulty of reconciling pairs of descriptions to determine whether they reflect compatible functions makes this analysis imprecise. Generally, the approach here is generous and should err on the side of detecting too few errors; it is usually more permissive than Ref. 5. mg463: Frasier *et al.*<sup>1</sup> and Koonin *et al.*<sup>2</sup> describe different aspects of function, but give the same gene name. The Ouzounis *et al.*<sup>3</sup> description is compatible with that from Koonin *et al.*<sup>2</sup>, but less specific. All three annotations are considered correct for this analysis. mg010: Frasier *et al.*<sup>1</sup> and Ouzounis *et al.*<sup>3</sup> agree that this is a DNA primase. Koonin *et al.*<sup>2</sup> use a different gene name and explicitly state that this is a truncated protein. Because of the common functional descriptions, all three are considered correct. However, if Koonin *et al.*<sup>2</sup> had been more explicit in indicating a functional difference, then their annotation would have been marked as conflicting. (Note that mg250 is also annotated as a DNA primase by all three groups.) mg225: the Ouzounis *et al.*<sup>3</sup> annotation of histidine permease is more specific than the Koonin *et al.*<sup>2</sup> description of amino acid permease. It may be that histidine permease is an (incorrect) overprediction of function, or it could be correct. The two annotations are considered consistent, and the decision of Frasier *et al.*<sup>1</sup> not to provide a function is not penalized. (b) Inconsistent annotations. mg302: lack of a functional assignment from Frasier *et al.*<sup>1</sup> is not penalized. The Koonin *et al.*<sup>2</sup> and Ouzounis *et al.*<sup>3</sup> annotations are wholly inconsistent. This leads to a conflict and a minimum error rate of 50%. Note that the assessment methodology also behaves correctly when two annotators provide different functions for a multi-functional enzyme: each of the annotators is half right and half wrong, and the assessment assigns a 50% error rate. mg448: Frasier *et al.*<sup>1</sup> and Ouzounis *et al.*<sup>3</sup> both describe the gene as *pilB*. The encoded protein is involved in pili formation, and its biochemical function is catalysis of methionine sulfoxide oxidation/reduction in proteins. The Koonin *et al.*<sup>2</sup> annotation, chaperone-like protein, could conceivably be compatible but this is not likely. Because of uncertainty regarding compatibility of the Koonin *et al.*<sup>2</sup> annotation and its qualification as putative, this set of annotations is right on the threshold of consideration. For this analysis, the Koonin *et al.*<sup>2</sup> annotation was considered to be in conflict with the others, giving a minimum error rate of 33%. mg085: all three groups provide contradictory functions: The function described by Frasier *et al.*<sup>1</sup> of HMG-CoA reductase is EC 1.1.1.34, while the NADH-ubiquinone oxidoreductase annotated by Ouzounis *et al.*<sup>3</sup> (nu6m\_marpo) is EC 1.6.5.3. Neither enzyme uses ATP or GTP, as specified by Koonin *et al.*<sup>2</sup>. The analysis assumes one is correct and marks two incorrect. Note: Ouzounis *et al.*<sup>3</sup> annotations equivalent to SWISS-PROT included in these examples are not included in the Table 1 analysis.

## Acknowledgements

A previous version of this analysis was performed at the MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK. M. Levitt, C. Chothia, B. Al-Lazikani and P. Koehl provided stimulating discussion.

## References

- 1 Frasier, C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397-403
- 2 Koonin, E.V. *et al.* (1996) Sequencing and analysis of bacterial genomes. *Curr. Biol.* 6, 404-416
- 3 Ouzounis, C. *et al.* (1996) Novelities from the complete genome of *Mycoplasma genitalium*. *Mol. Microbiol.* 20, 898-900
- 4 Doerks, T. *et al.* (1998) Protein annotation: detective work for function prediction. *Trends Genet.* 14, 248-250
- 5 Galperin, M.Y. and Koonin, E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol.* 1, 7
- 6 Smith, T.F. and Zhang, X. (1997) The challenges of genome sequence annotation or 'The devil is in the details'. *Nat. Biotechnol.* 15, 1222-1223
- 7 Bork, P. *et al.* (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283, 707-725
- 8 Bork, P. and Bairoch, A. (1996) Go hunting in sequence databases but watch out for the traps. *Trends Genet.* 12, 425-427
- 9 Brenner, S.E. *et al.* (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6073-6078
- 10 Altschul, S.F. *et al.* (1994) Issues in searching molecular sequence databases. *Nat. Genet.* 6, 119-129
- 11 Chothia, C. (1992) Proteins. One thousand families for the molecular biologist. *Nature* 357, 543-544
- 12 Brenner, S.E. *et al.* (1997) Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* 7, 369-376
- 13 Smith, T.F. (1998) Functional genomics - bioinformatics is ready for the challenge. *Trends Genet.* 14, 291-329
- 14 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631-637

**TABLE 1. *M. genitalium* annotations, conflicts and error rates**

No. groups annotating gene	No. genes	Annotations per group <sup>a</sup>			Total annotations	No. conflicts	Minimum error rate
		Frasier <i>et al.</i> <sup>1</sup>	Koonin <i>et al.</i> <sup>2</sup>	Ouzounis <i>et al.</i> <sup>3</sup>			
0	33	—	—	—	—	N/A	N/A
1 <sup>a</sup>	95	14	15	66	95	N/A	N/A
2	318	279	317	40	636	45	7%
3	22	22	22	22	66	10	15%
Sum (2+3)	340	301	339	62	702	55	8%

Summary of annotations made by each group (Fig. 1), minimal number of conflicting annotations (see Fig. 2), the resulting minimal fraction of annotations that are erroneous.

<sup>a</sup>Frasier *et al.*<sup>1</sup> data from <http://www.tigr.org/tdb/mdb/mgdb/mgdb.html>, Koonin *et al.*<sup>2</sup> data from [http://www.nlm.nih.gov/Complete\\_Genomes/Mgen](http://www.nlm.nih.gov/Complete_Genomes/Mgen), Ouzounis *et al.*<sup>3</sup> data from <http://www.embl-heidelberg.de/~gene/mycogen.new.html>. Instances where Ouzounis *et al.*<sup>3</sup> reported SWISS-PROT annotation of the same gene were removed to avoid duplication with Frasier *et al.*<sup>1</sup> entries. However, even if all of these 300 annotations are included, the minimum annotation error rate drops only to 6%. All annotations were collected in 1996, shortly after the genome was released. No comparative analysis is possible when only one group made an annotation.

## **WISP genes are members of the connective tissue growth factor family that are up-regulated in Wnt-1-transformed cells and aberrantly expressed in human colon tumors**

DIANE PENNICA\*<sup>†</sup>, TODD A. SWANSON\*, JAMES W. WELSH\*, MARGARET A. ROY<sup>‡</sup>, DAVID A. LAWRENCE\*, JAMES LEE<sup>‡</sup>, JENNIFER BRUSH<sup>‡</sup>, LISA A. TANEYHILL<sup>§</sup>, BETHANNE DEUEL<sup>‡</sup>, MICHAEL LEW<sup>¶</sup>, COLIN WATANABE<sup>¶</sup>, ROBERT L. COHEN\*, MONA F. MELHEM\*\*, GENE G. FINLEY\*\*, PHIL QUIRKE<sup>††</sup>, AUDREY D. GODDARD<sup>‡</sup>, KENNETH J. HILLAN<sup>¶</sup>, AUSTIN L. GURNEY<sup>‡</sup>, DAVID BOTSTEIN<sup>‡,‡‡</sup>, AND ARNOLD J. LEVINE<sup>§</sup>

Departments of \*Molecular Oncology, <sup>‡</sup>Molecular Biology, <sup>§</sup>Scientific Computing, and <sup>¶</sup>Pathology, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080; \*\*University of Pittsburgh School of Medicine, Veterans Administration Medical Center, Pittsburgh, PA 15240; <sup>††</sup>University of Leeds, Leeds, LS29JT United Kingdom; <sup>‡‡</sup>Department of Genetics, Stanford University, Palo Alto, CA 94305; and <sup>§</sup>Department of Molecular Biology, Princeton University, Princeton, NJ 08544

Contributed by David Botstein and Arnold J. Levine, October 21, 1998

**ABSTRACT** Wnt family members are critical to many developmental processes, and components of the Wnt signaling pathway have been linked to tumorigenesis in familial and sporadic colon carcinomas. Here we report the identification of two genes, *WISP-1* and *WISP-2*, that are up-regulated in the mouse mammary epithelial cell line C57MG transformed by Wnt-1, but not by Wnt-4. Together with a third related gene, *WISP-3*, these proteins define a subfamily of the connective tissue growth factor family. Two distinct systems demonstrated *WISP* induction to be associated with the expression of Wnt-1. These included (i) C57MG cells infected with a Wnt-1 retroviral vector or expressing Wnt-1 under the control of a tetracycline repressible promoter, and (ii) Wnt-1 transgenic mice. The *WISP-1* gene was localized to human chromosome 8q24.1-8q24.3. *WISP-1* genomic DNA was amplified in colon cancer cell lines and in human colon tumors and its RNA overexpressed (2- to >30-fold) in 84% of the tumors examined compared with patient-matched normal mucosa. *WISP-3* mapped to chromosome 6q22-6q23 and also was overexpressed (4- to >40-fold) in 63% of the colon tumors analyzed. In contrast, *WISP-2* mapped to human chromosome 20q12-20q13 and its DNA was amplified, but RNA expression was reduced (2- to >30-fold) in 79% of the tumors. These results suggest that the *WISP* genes may be downstream of Wnt-1 signaling and that aberrant levels of *WISP* expression in colon cancer may play a role in colon tumorigenesis.

Wnt-1 is a member of an expanding family of cysteine-rich, glycosylated signaling proteins that mediate diverse developmental processes such as the control of cell proliferation, adhesion, cell polarity, and the establishment of cell fates (1, 2). Wnt-1 originally was identified as an oncogene activated by the insertion of mouse mammary tumor virus in virus-induced mammary adenocarcinomas (3, 4). Although Wnt-1 is not expressed in the normal mammary gland, expression of Wnt-1 in transgenic mice causes mammary tumors (5).

In mammalian cells, Wnt family members initiate signaling by binding to the seven-transmembrane spanning Frizzled receptors and recruiting the cytoplasmic protein Dishevelled (Dsh) to the cell membrane (1, 2, 6). Dsh then inhibits the kinase activity of the normally constitutively active glycogen synthase kinase-3 $\beta$  (GSK-3 $\beta$ ) resulting in an increase in  $\beta$ -catenin levels. Stabilized  $\beta$ -catenin interacts with the transcription factor TCF/Lef1, forming a complex that appears in

the nucleus and binds TCF/Lef1 target DNA elements to activate transcription (7, 8). Other experiments suggest that the adenomatous polyposis coli (APC) tumor suppressor gene also plays an important role in Wnt signaling by regulating  $\beta$ -catenin levels (9). APC is phosphorylated by GSK-3 $\beta$ , binds to  $\beta$ -catenin, and facilitates its degradation. Mutations in either APC or  $\beta$ -catenin have been associated with colon carcinomas and melanomas, suggesting these mutations contribute to the development of these types of cancer, implicating the Wnt pathway in tumorigenesis (1).

Although much has been learned about the Wnt signaling pathway over the past several years, only a few of the transcriptionally activated downstream components activated by Wnt have been characterized. Those that have been described cannot account for all of the diverse functions attributed to Wnt signaling. Among the candidate Wnt target genes are those encoding the nodal-related 3 gene, *Xnr3*, a member of the transforming growth factor (TGF)- $\beta$  superfamily, and the homeobox genes, *engrailed*, *gooseoid*, *twin* (*Xtwn*), and *siamois* (2). A recent report also identifies *c-myc* as a target gene of the Wnt signaling pathway (10).

To identify additional downstream genes in the Wnt signaling pathway that are relevant to the transformed cell phenotype, we used a PCR-based cDNA subtraction strategy, suppression subtractive hybridization (SSH) (11), using RNA isolated from C57MG mouse mammary epithelial cells and C57MG cells stably transformed by a Wnt-1 retrovirus. Overexpression of Wnt-1 in this cell line is sufficient to induce a partially transformed phenotype, characterized by elongated and refractile cells that lose contact inhibition and form a multilayered array (12, 13). We reasoned that genes differentially expressed between these two cell lines might contribute to the transformed phenotype.

In this paper, we describe the cloning and characterization of two genes up-regulated in Wnt-1 transformed cells, *WISP-1* and *WISP-2*, and a third related gene, *WISP-3*. The *WISP* genes are members of the CCN family of growth factors, which includes connective tissue growth factor (CTGF), Cyr61, and *nov*, a family not previously linked to Wnt signaling.

### **MATERIALS AND METHODS**

**SSH.** SSH was performed by using the PCR-Select cDNA Subtraction Kit (CLONTECH). Tester double-stranded

Abbreviations: TGF, transforming growth factor; CTGF, connective tissue growth factor; SSH, suppression subtractive hybridization; WVC, von Willebrand factor type C module.

Data deposition: The sequences reported in this paper have been deposited in the Genbank database (accession nos. AF100777, AF100778, AF100779, AF100780, and AF100781).

<sup>†</sup>To whom reprint requests should be addressed. e-mail: diane@gene.com.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/9514717-6\$2.00/0 PNAS is available online at www.pnas.org.

cDNA was synthesized from 2  $\mu$ g of poly(A)<sup>+</sup> RNA isolated from the C57MG/Wnt-1 cell line and driver cDNA from 2  $\mu$ g of poly(A)<sup>+</sup> RNA from the parent C57MG cells. The subtracted cDNA library was subcloned into a pGEM-T vector for further analysis.

**cDNA Library Screening.** Clones encoding full-length mouse *WISP-1* were isolated by screening a  $\lambda$ gt10 mouse embryo cDNA library (CLONTECH) with a 70-bp probe from the original partial clone 568 sequence corresponding to amino acids 128–169. Clones encoding full-length human *WISP-1* were isolated by screening  $\lambda$ gt10 lung and fetal kidney cDNA libraries with the same probe at low stringency. Clones encoding full-length mouse and human *WISP-2* were isolated by screening a C57MG/Wnt-1 or human fetal lung cDNA library with a probe corresponding to nucleotides 1463–1512. Full-length cDNAs encoding *WISP-3* were cloned from human bone marrow and fetal kidney libraries.

**Expression of Human *WISP* RNA.** PCR amplification of first-strand cDNA was performed with human Multiple Tissue cDNA panels (CLONTECH) and 300  $\mu$ M of each dNTP at 94°C for 1 sec, 62°C for 30 sec, 72°C for 1 min, for 22–32 cycles. *WISP* and glyceraldehyde-3-phosphate dehydrogenase primer sequences are available on request.

**In Situ Hybridization.** <sup>33</sup>P-labeled sense and antisense riboprobes were transcribed from an 897-bp PCR product corresponding to nucleotides 601–1440 of mouse *WISP-1* or a 294-bp PCR product corresponding to nucleotides 82–375 of mouse *WISP-2*. All tissues were processed as described (40).

**Radiation Hybrid Mapping.** Genomic DNA from each hybrid in the Stanford G3 and Genebridge4 Radiation Hybrid Panels (Research Genetics, Huntsville, AL) and human and hamster control DNAs were PCR-amplified, and the results were submitted to the Stanford or Massachusetts Institute of Technology web servers.

**Cell Lines, Tumors, and Mucosa Specimens.** Tissue specimens were obtained from the Department of Pathology (University of Pittsburgh) for patients undergoing colon resection and from the University of Leeds, United Kingdom. Genomic DNA was isolated (Qiagen) from the pooled blood of 10 normal human donors, surgical specimens, and the following ATCC human cell lines: SW480, COLO 320DM, HT-29, WiDr, and SW403 (colon adenocarcinomas), SW620 (lymph node metastasis, colon adenocarcinoma), HCT 116 (colon carcinoma), SK-CO-1 (colon adenocarcinoma, ascites), and HM7 (a variant of ATCC colon adenocarcinoma cell line LS 174T). DNA concentration was determined by using Hoechst dye 33258 intercalation fluorimetry. Total RNA was prepared by homogenization in 7 M GuSCN followed by centrifugation over CsCl cushions or prepared by using RNeasy.

**Gene Amplification and RNA Expression Analysis.** Relative gene amplification and RNA expression of *WISPs* and *c-myc* in the cell lines, colorectal tumors, and normal mucosa were determined by quantitative PCR. Gene-specific primers and fluorogenic probes (sequences available on request) were designed and used to amplify and quantitate the genes. The relative gene copy number was derived by using the formula  $2^{\Delta\Delta Ct}$  where  $\Delta Ct$  represents the difference in amplification cycles required to detect the *WISP* genes in peripheral blood lymphocyte DNA compared with colon tumor DNA or colon tumor RNA compared with normal mucosal RNA. The  $\Delta$ -method was used for calculation of the SE of the gene copy number or RNA expression level. The *WISP*-specific signal was normalized to that of the glyceraldehyde-3-phosphate dehydrogenase housekeeping gene. All TaqMan assay reagents were obtained from Perkin-Elmer Applied Biosystems.

## RESULTS

**Isolation of *WISP-1* and *WISP-2* by SSH.** To identify Wnt-1-inducible genes, we used the technique of SSH using the

mouse mammary epithelial cell line C57MG and C57MG cells that stably express Wnt-1 (11). Candidate differentially expressed cDNAs (1,384 total) were sequenced. Thirty-nine percent of the sequences matched known genes or homologues, 32% matched expressed sequence tags, and 29% had no match. To confirm that the transcript was differentially expressed, semiquantitative reverse transcription-PCR and Northern analysis were performed by using mRNA from the C57MG and C57MG/Wnt-1 cells.

Two of the cDNAs, *WISP-1* and *WISP-2*, were differentially expressed, being induced in the C57MG/Wnt-1 cell line, but not in the parent C57MG cells or C57MG cells overexpressing Wnt-4 (Fig. 1A and B). Wnt-4, unlike Wnt-1, does not induce the morphological transformation of C57MG cells and has no effect on  $\beta$ -catenin levels (13, 14). Expression of *WISP-1* was up-regulated approximately 3-fold in the C57MG/Wnt-1 cell line and *WISP-2* by approximately 5-fold by both Northern analysis and reverse transcription-PCR.

An independent, but similar, system was used to examine *WISP* expression after Wnt-1 induction. C57MG cells expressing the *Wnt-1* gene under the control of a tetracycline-repressible promoter produce low amounts of Wnt-1 in the repressed state but show a strong induction of Wnt-1 mRNA and protein within 24 hr after tetracycline removal (8). The levels of Wnt-1 and *WISP* RNA isolated from these cells at various times after tetracycline removal were assessed by quantitative PCR. Strong induction of Wnt-1 mRNA was seen as early as 10 hr after tetracycline removal. Induction of *WISP* mRNA (2- to 6-fold) was seen at 48 and 72 hr (data not shown). These data support our previous observations that show that *WISP* induction is correlated with Wnt-1 expression. Because the induction is slow, occurring after approximately 48 hr, the induction of *WISPs* may be an indirect response to Wnt-1 signaling.

cDNA clones of human *WISP-1* were isolated and the sequence compared with mouse *WISP-1*. The cDNA sequences of mouse and human *WISP-1* were 1,766 and 2,830 bp in length, respectively, and encode proteins of 367 aa, with predicted relative molecular masses of  $\approx$ 40,000 ( $M_r$  40 K). Both have hydrophobic N-terminal signal sequences, 38 conserved cysteine residues, and four potential N-linked glycosylation sites and are 84% identical (Fig. 2A).

Full-length cDNA clones of mouse and human *WISP-2* were 1,734 and 1,293 bp in length, respectively, and encode proteins of 251 and 250 aa, respectively, with predicted relative molecular masses of  $\approx$ 27,000 ( $M_r$  27 K) (Fig. 2B). Mouse and human *WISP-2* are 73% identical. Human *WISP-2* has no potential N-linked glycosylation sites, and mouse *WISP-2* has one at

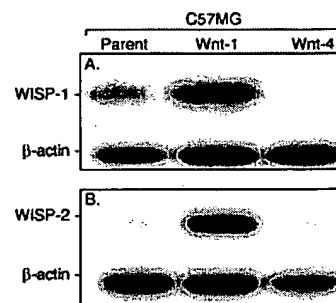


FIG. 1. *WISP-1* and *WISP-2* are induced by Wnt-1, but not Wnt-4, expression in C57MG cells. Northern analysis of *WISP-1* (A) and *WISP-2* (B) expression in C57MG, C57MG/Wnt-1, and C57MG/Wnt-4 cells. Poly(A)<sup>+</sup> RNA (2  $\mu$ g) was subjected to Northern blot analysis and hybridized with a 70-bp mouse *WISP-1*-specific probe (amino acids 278–300) or a 190-bp *WISP-2*-specific probe (nucleotides 1438–1627) in the 3' untranslated region. Blots were rehybridized with human  $\beta$ -actin probe.



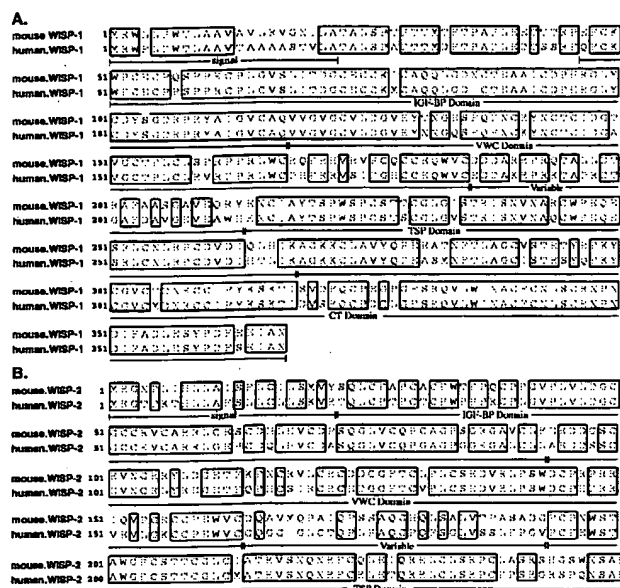


FIG. 2. Encoded amino acid sequence alignment of mouse and human *WISP-1* (A) and mouse and human *WISP-2* (B). The potential signal sequence, insulin-like growth factor-binding protein (IGF-BP), VWC, thrombospondin (TSP), and C-terminal (CT) domains are underlined.

position 197. *WISP-2* has 28 cysteine residues that are conserved among the 38 cysteines found in *WISP-1*.

**Identification of *WISP-3*.** To search for related proteins, we screened expressed sequence tag (EST) databases with the *WISP-1* protein sequence and identified several ESTs as potentially related sequences. We identified a homologous protein that we have called *WISP-3*. A full-length human *WISP-3* cDNA of 1,371 bp was isolated corresponding to those ESTs that encode a 354-aa protein with a predicted molecular mass of 39,293. *WISP-3* has two potential N-linked glycosylation sites and 36 cysteine residues. An alignment of the three human *WISP* proteins shows that *WISP-1* and *WISP-3* are the most similar (42% identity), whereas *WISP-2* has 37% identity with *WISP-1* and 32% identity with *WISP-3* (Fig. 3A).

***WISPs* Are Homologous to the CTGF Family of Proteins.** Human *WISP-1*, *WISP-2*, and *WISP-3* are novel sequences; however, mouse *WISP-1* is the same as the recently identified *Elm1* gene. *Elm1* is expressed in low, but not high, metastatic mouse melanoma cells, and suppresses the *in vivo* growth and metastatic potential of K-1735 mouse melanoma cells (15). Human and mouse *WISP-2* are homologous to the recently described rat gene, *rCop-1* (16). Significant homology (36–44%) was seen to the CCN family of growth factors. This family includes three members, CTGF, Cyr61, and the protooncogene *nov*. CTGF is a chemotactic and mitogenic factor for fibroblasts that is implicated in wound healing and fibrotic disorders and is induced by TGF- $\beta$  (17). Cyr61 is an extracellular matrix signaling molecule that promotes cell adhesion, proliferation, migration, angiogenesis, and tumor growth (18, 19). *nov* (nephroblastoma overexpressed) is an immediate early gene associated with quiescence and found altered in Wilms tumors (20). The proteins of the CCN family share functional, but not sequence, similarity to Wnt-1. All are secreted, cysteine-rich heparin binding glycoproteins that associate with the cell surface and extracellular matrix.

*WISP* proteins exhibit the modular architecture of the CCN family, characterized by four conserved cysteine-rich domains (Fig. 3B) (21). The N-terminal domain, which includes the first 12 cysteine residues, contains a consensus sequence (GCGC-CXXC) conserved in most insulin-like growth factor (IGF)-

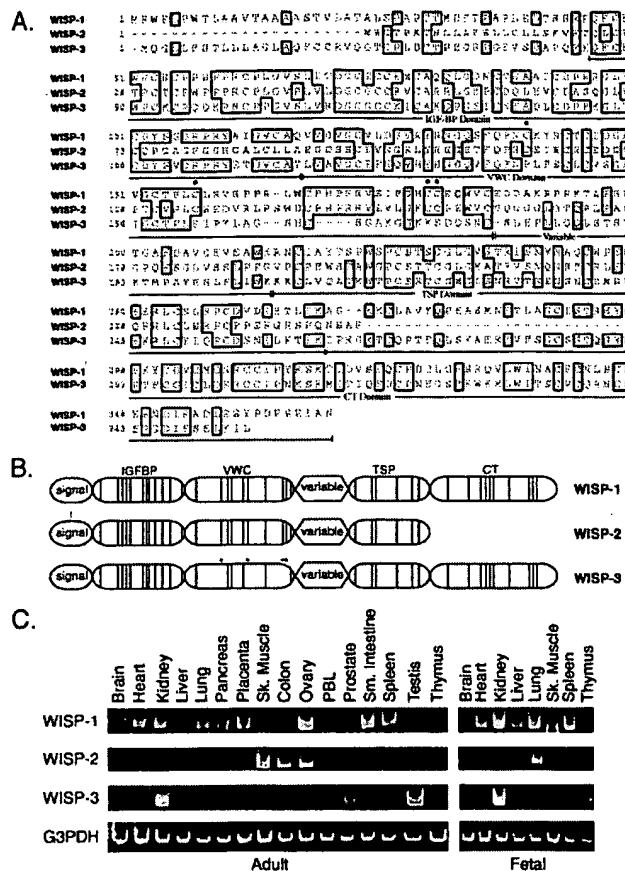


FIG. 3. (A) Encoded amino acid sequence alignment of human *WISPs*. The cysteine residues of *WISP-1* and *WISP-2* that are not present in *WISP-3* are indicated with a dot. (B) Schematic representation of the *WISP* proteins showing the domain structure and cysteine residues (vertical lines). The four cysteine residues in the VWC domain that are absent in *WISP-3* are indicated with a dot. (C) Expression of *WISP* mRNA in human tissues. PCR was performed on human multiple-tissue cDNA panels (CLONTECH) from the indicated adult and fetal tissues.

binding proteins (BP). This sequence is conserved in *WISP-2* and *WISP-3*, whereas *WISP-1* has a glutamine in the third position instead of a glycine. CTGF recently has been shown to specifically bind IGF (22) and a truncated *nov* protein lacking the IGF-BP domain is oncogenic (23). The von Willebrand factor type C module (VWC), also found in certain collagens and mucins, covers the next 10 cysteine residues, and is thought to participate in protein complex formation and oligomerization (24). The VWC domain of *WISP-3* differs from all CCN family members described previously, in that it contains only six of the 10 cysteine residues (Fig. 3A and B). A short variable region follows the VWC domain. The third module, the thrombospondin (TSP) domain is involved in binding to sulfated glycoconjugates and contains six cysteine residues and a conserved WSXCSXCG motif first identified in thrombospondin (25). The C-terminal (CT) module containing the remaining 10 cysteines is thought to be involved in dimerization and receptor binding (26). The CT domain is present in all CCN family members described to date but is absent in *WISP-2* (Fig. 3A and B). The existence of a putative signal sequence and the absence of a transmembrane domain suggest that *WISPs* are secreted proteins, an observation supported by an analysis of their expression and secretion from mammalian cell and baculovirus cultures (data not shown).

**Expression of *WISP* mRNA in Human Tissues.** Tissue-specific expression of human *WISPs* was characterized by PCR



analysis on adult and fetal multiple tissue cDNA panels. *WISP-1* expression was seen in the adult heart, kidney, lung, pancreas, placenta, ovary, small intestine, and spleen (Fig. 3C). Little or no expression was detected in the brain, liver, skeletal muscle, colon, peripheral blood leukocytes, prostate, testis, or thymus. *WISP-2* had a more restricted tissue expression and was detected in adult skeletal muscle, colon, ovary, and fetal lung. Predominant expression of *WISP-3* was seen in adult kidney and testis and fetal kidney. Lower levels of *WISP-3* expression were detected in placenta, ovary, prostate, and small intestine.

**In Situ Localization of *WISP-1* and *WISP-2*.** Expression of *WISP-1* and *WISP-2* was assessed by *in situ* hybridization in mammary tumors from Wnt-1 transgenic mice. Strong expression of *WISP-1* was observed in stromal fibroblasts lying within the fibrovascular tumor stroma (Fig. 4 A–D). However, low-level *WISP-1* expression also was observed focally within tumor cells (data not shown). No expression was observed in normal breast. Like *WISP-1*, *WISP-2* expression also was seen in the tumor stroma in breast tumors from Wnt-1 transgenic animals (Fig. 4 E–H). However, *WISP-2* expression in the stroma was in spindle-shaped cells adjacent to capillary vessels, whereas

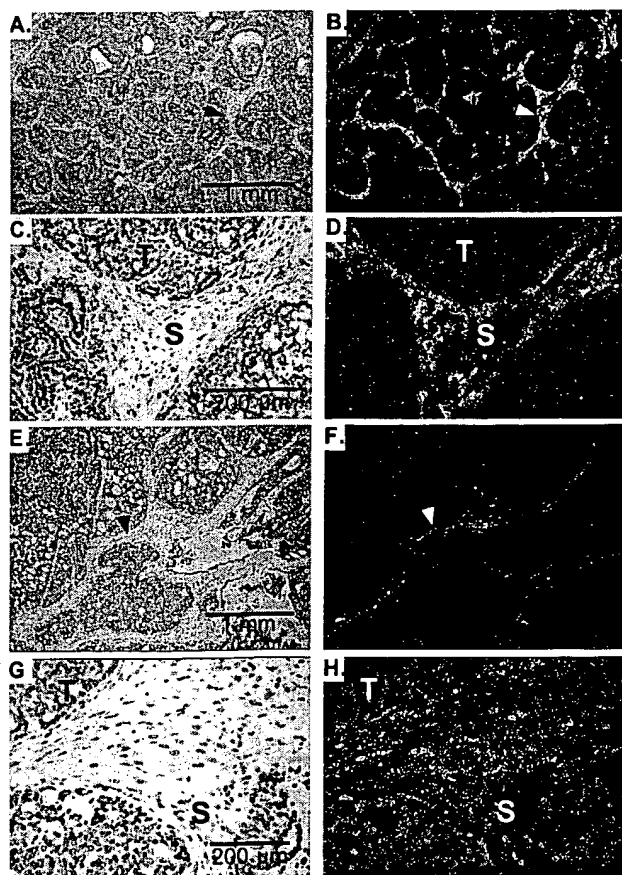


FIG. 4. (A, C, E, and G) Representative hematoxylin/eosin-stained images from breast tumors in Wnt-1 transgenic mice. The corresponding dark-field images showing *WISP-1* expression are shown in B and D. The tumor is a moderately well-differentiated adenocarcinoma showing evidence of adenoid cystic change. At low power (A and B), expression of *WISP-1* is seen in the delicate branching fibrovascular tumor stroma (arrowhead). At higher magnification, expression is seen in the stromal(s) fibroblasts (C and D), and tumor cells are negative. Focal expression of *WISP-1*, however, was observed in tumor cells in some areas. Images of *WISP-2* expression are shown in E–H. At low power (E and F), expression of *WISP-2* is seen in cells lying within the fibrovascular tumor stroma. At higher magnification, these cells appeared to be adjacent to capillary vessels whereas tumor cells are negative (G and H).

the predominant cell type expressing *WISP-1* was the stromal fibroblasts.

**Chromosome Localization of the *WISP* Genes.** The chromosomal location of the human *WISP* genes was determined by radiation hybrid mapping panels. *WISP-1* is approximately 3.48 cR from the meiotic marker AFM259xc5 [logarithm of odds (lod) score 16.31] on chromosome 8q24.1 to 8q24.3, in the same region as the human locus of the *novH* family member (27) and roughly 4 Mbs distal to *c-myc* (28). Preliminary fine mapping indicates that *WISP-1* is located near D8S1712 STS. *WISP-2* is linked to the marker SHGC-33922 (lod = 1,000) on chromosome 20q12–20q13.1. Human *WISP-3* mapped to chromosome 6q22–6q23 and is linked to the marker AFM211ze5 (lod = 1,000). *WISP-3* is approximately 18 Mbs proximal to CTGF and 23 Mbs proximal to the human cellular oncogene *MYB* (27, 29).

**Amplification and Aberrant Expression of *WISPs* in Human Colon Tumors.** Amplification of protooncogenes is seen in many human tumors and has etiological and prognostic significance. For example, in a variety of tumor types, *c-myc* amplification has been associated with malignant progression and poor prognosis (30). Because *WISP-1* resides in the same general chromosomal location (8q24) as *c-myc*, we asked whether it was a target of gene amplification, and, if so, whether this amplification was independent of the *c-myc* locus. Genomic DNA from human colon cancer cell lines was assessed by quantitative PCR and Southern blot analysis. (Fig. 5 A and B). Both methods detected similar degrees of *WISP-1* amplification. Most cell lines showed significant (2- to 4-fold) amplification, with the HT-29 and WiDr cell lines demonstrating an 8-fold increase. Significantly, the pattern of amplification observed did not correlate with that observed for *c-myc*, indicating that the *c-myc* gene is not part of the amplicon that involves the *WISP-1* locus.

We next examined whether the *WISP* genes were amplified in a panel of 25 primary human colon adenocarcinomas. The relative *WISP* gene copy number in each colon tumor DNA was compared with pooled normal DNA from 10 donors by quantitative PCR (Fig. 6). The copy number of *WISP-1* and *WISP-2* was significantly greater than one, approximately 2-fold for *WISP-1* in about 60% of the tumors and 2- to 4-fold for *WISP-2* in 92% of the tumors ( $P < 0.001$  for each). The copy number for *WISP-3* was indistinguishable from one ( $P = 0.166$ ). In addition, the copy number of *WISP-2* was significantly higher than that of *WISP-1* ( $P < 0.001$ ).

The levels of *WISP* transcripts in RNA isolated from 19 adenocarcinomas and their matched normal mucosa were

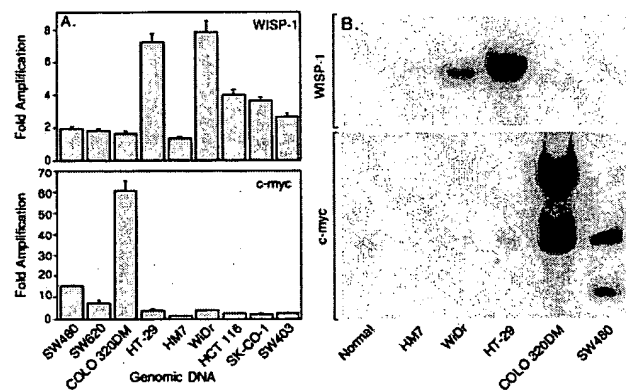


FIG. 5. Amplification of *WISP-1* genomic DNA in colon cancer cell lines. (A) Amplification in cell line DNA was determined by quantitative PCR. (B) Southern blots containing genomic DNA (10  $\mu$ g) digested with *EcoRI* (*WISP-1*) or *XbaI* (*c-myc*) were hybridized with a 100-bp human *WISP-1* probe (amino acids 186–219) or a human *c-myc* probe (located at bp 1901–2000). The *WISP* and *myc* genes are detected in normal human genomic DNA after a longer film exposure.

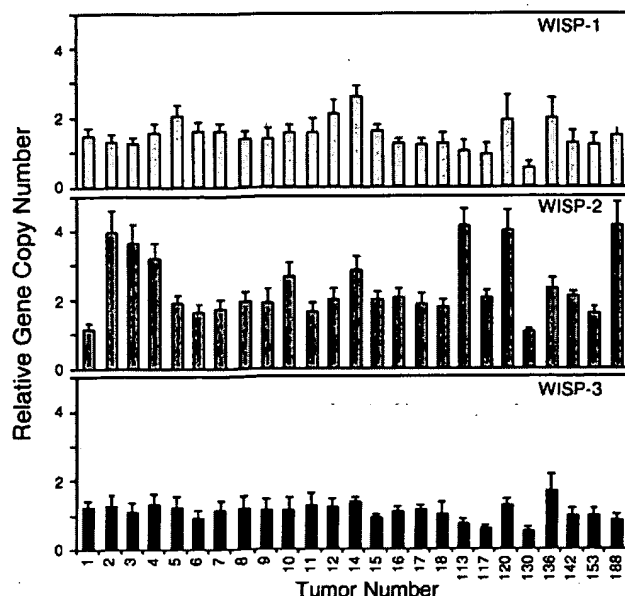


FIG. 6. Genomic amplification of *WISP* genes in human colon tumors. The relative gene copy number of the *WISP* genes in 25 adenocarcinomas was assayed by quantitative PCR, by comparing DNA from primary human tumors with pooled DNA from 10 healthy donors. The data are means  $\pm$  SEM from one experiment done in triplicate. The experiment was repeated at least three times.

assessed by quantitative PCR (Fig. 7). The level of *WISP-1* RNA present in tumor tissue varied but was significantly increased (2- to >25-fold) in 84% (16/19) of the human colon tumors examined compared with normal adjacent mucosa. Four of 19 tumors showed greater than 10-fold overexpression. In contrast, in 79% (15/19) of the tumors examined, *WISP-2* RNA expression was significantly lower in the tumor than the mucosa. Similar to *WISP-1*, *WISP-3* RNA was overexpressed in 63% (12/19) of the colon tumors compared with the normal

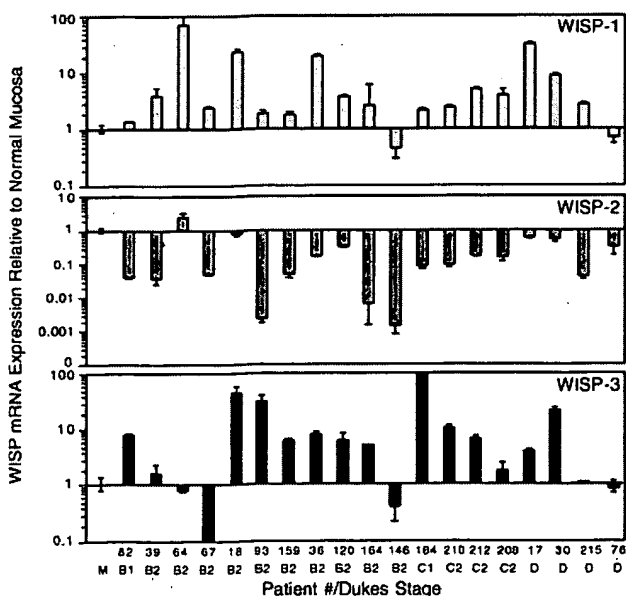


FIG. 7. *WISP* RNA expression in primary human colon tumors relative to expression in normal mucosa from the same patient. Expression of *WISP* mRNA in 19 adenocarcinomas was assayed by quantitative PCR. The Dukes stage of the tumor is listed under the sample number. The data are means  $\pm$  SEM from one experiment done in triplicate. The experiment was repeated at least twice.

mucosa. The amount of overexpression of *WISP-3* ranged from 4- to >40-fold.

## DISCUSSION

One approach to understanding the molecular basis of cancer is to identify differences in gene expression between cancer cells and normal cells. Strategies based on assumptions that steady-state mRNA levels will differ between normal and malignant cells have been used to clone differentially expressed genes (31). We have used a PCR-based selection strategy, SSH, to identify genes selectively expressed in C57MG mouse mammary epithelial cells transformed by Wnt-1.

Three of the genes isolated, *WISP-1*, *WISP-2*, and *WISP-3*, are members of the CCN family of growth factors, which includes CTGF, Cyr61, and *nov*, a family not previously linked to Wnt signaling.

Two independent experimental systems demonstrated that *WISP* induction was associated with the expression of Wnt-1. The first was C57MG cells infected with a Wnt-1 retroviral vector or C57MG cells expressing Wnt-1 under the control of a tetracycline-repressible promoter, and the second was in Wnt-1 transgenic mice, where breast tissue expresses Wnt-1, whereas normal breast tissue does not. No *WISP* RNA expression was detected in mammary tumors induced by polyoma virus middle T antigen (data not shown). These data suggest a link between Wnt-1 and *WISPs* in that in these two situations, *WISP* induction was correlated with Wnt-1 expression.

It is not clear whether the *WISPs* are directly or indirectly induced by the downstream components of the Wnt-1 signaling pathway (i.e.,  $\beta$ -catenin-TCF-1/Lef1). The increased levels of *WISP* RNA were measured in Wnt-1-transformed cells, hours or days after Wnt-1 transformation. Thus, *WISP* expression could result from Wnt-1 signaling directly through  $\beta$ -catenin transcription factor regulation or alternatively through Wnt-1 signaling turning on a transcription factor, which in turn regulates *WISPs*.

The *WISPs* define an additional subfamily of the CCN family of growth factors. One striking difference observed in the protein sequence of *WISP-2* is the absence of a CT domain, which is present in CTGF, Cyr61, *nov*, *WISP-1*, and *WISP-3*. This domain is thought to be involved in receptor binding and dimerization. Growth factors, such as TGF- $\beta$ , platelet-derived growth factor, and nerve growth factor, which contain a cystine knot motif exist as dimers (32). It is tempting to speculate that *WISP-1* and *WISP-3* may exist as dimers, whereas *WISP-2* exists as a monomer. If the CT domain is also important for receptor binding, *WISP-2* may bind its receptor through a different region of the molecule than the other CCN family members. No specific receptors have been identified for CTGF or *nov*. A recent report has shown that integrin  $\alpha_v\beta_3$  serves as an adhesion receptor for Cyr61 (33).

The strong expression of *WISP-1* and *WISP-2* in cells lying within the fibrovascular tumor stroma in breast tumors from Wnt-1 transgenic animals is consistent with previous observations that transcripts for the related CTGF gene are primarily expressed in the fibrous stroma of mammary tumors (34). Epithelial cells are thought to control the proliferation of connective tissue stroma in mammary tumors by a cascade of growth factor signals similar to that controlling connective tissue formation during wound repair. It has been proposed that mammary tumor cells or inflammatory cells at the tumor interstitial interface secrete TGF- $\beta$ 1, which is the stimulus for stromal proliferation (34). TGF- $\beta$ 1 is secreted by a large percentage of malignant breast tumors and may be one of the growth factors that stimulates the production of CTGF and *WISPs* in the stroma.

It was of interest that *WISP-1* and *WISP-2* expression was observed in the stromal cells that surrounded the tumor cells

(epithelial cells) in the Wnt-1 transgenic mouse sections of breast tissue. This finding suggests that paracrine signaling could occur in which the stromal cells could supply WISP-1 and WISP-2 to regulate tumor cell growth on the WISP extracellular matrix. Stromal cell-derived factors in the extracellular matrix have been postulated to play a role in tumor cell migration and proliferation (35). The localization of *WISP-1* and *WISP-2* in the stromal cells of breast tumors supports this paracrine model.

An analysis of *WISP-1* gene amplification and expression in human colon tumors showed a correlation between DNA amplification and overexpression, whereas overexpression of *WISP-3* RNA was seen in the absence of DNA amplification. In contrast, *WISP-2* DNA was amplified in the colon tumors, but its mRNA expression was significantly reduced in the majority of tumors compared with the expression in normal colonic mucosa from the same patient. The gene for human *WISP-2* was localized to chromosome 20q12–20q13, at a region frequently amplified and associated with poor prognosis in node negative breast cancer and many colon cancers, suggesting the existence of one or more oncogenes at this locus (36–38). Because the center of the 20q13 amplicon has not yet been identified, it is possible that the apparent amplification observed for *WISP-2* may be caused by another gene in this amplicon.

A recent manuscript on *rCop-1*, the rat orthologue of *WISP-2*, describes the loss of expression of this gene after cell transformation, suggesting it may be a negative regulator of growth in cell lines (16). Although the mechanism by which *WISP-2* RNA expression is down-regulated during malignant transformation is unknown, the reduced expression of *WISP-2* in colon tumors and cell lines suggests that it may function as a tumor suppressor. These results show that the *WISP* genes are aberrantly expressed in colon cancer and suggest that their altered expression may confer selective growth advantage to the tumor.

Members of the Wnt signaling pathway have been implicated in the pathogenesis of colon cancer, breast cancer, and melanoma, including the tumor suppressor gene adenomatous polyposis coli and  $\beta$ -catenin (39). Mutations in specific regions of either gene can cause the stabilization and accumulation of cytoplasmic  $\beta$ -catenin, which presumably contributes to human carcinogenesis through the activation of target genes such as the *WISPs*. Although the mechanism by which Wnt-1 transforms cells and induces tumorigenesis is unknown, the identification of *WISPs* as genes that may be regulated downstream of Wnt-1 in C57MG cells suggests they could be important mediators of Wnt-1 transformation. The amplification and altered expression patterns of the *WISPs* in human colon tumors may indicate an important role for these genes in tumor development.

We thank the DNA synthesis group for oligonucleotide synthesis, T. Baker for technical assistance, P. Dowd for radiation hybrid mapping, K. Willert and R. Nusse for the tet-repressible C57MG/Wnt-1 cells, V. Dixit for discussions, and D. Wood and A. Bruce for artwork.

- Cadigan, K. M. & Nusse, R. (1997) *Genes Dev.* **11**, 3286–3305.
- Dale, T. C. (1998) *Biochem. J.* **329**, 209–223.
- Nusse, R. & Varmus, H. E. (1982) *Cell* **31**, 99–109.
- van Ooyen, A. & Nusse, R. (1984) *Cell* **39**, 233–240.
- Tsukamoto, A. S., Grosschedl, R., Guzman, R. C., Parslow, T. & Varmus, H. E. (1988) *Cell* **55**, 619–625.
- Brown, J. D. & Moon, R. T. (1998) *Curr. Opin. Cell Biol.* **10**, 182–187.
- Molenaar, M., van de Wetering, M., Oosterwegel, M., Peterson-Maduro, J., Godsave, S., Korinek, V., Roose, J., Destree, O. & Clevers, H. (1996) *Cell* **86**, 391–399.
- Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O. & Clevers, H. (1998) *Mol. Cell Biol.* **18**, 1248–1256.
- Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B. & Polakis, P. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 3046–3050.
- He, T. C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B. & Kinzler, K. W. (1998) *Science* **281**, 1509–1512.
- Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. & Siebert, P. D. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 6025–6030.
- Brown, A. M., Wildin, R. S., Prendergast, T. J. & Varmus, H. E. (1986) *Cell* **46**, 1001–1009.
- Wong, G. T., Gavin, B. J. & McMahon, A. P. (1994) *Mol. Cell Biol.* **14**, 6278–6286.
- Shimizu, H., Julius, M. A., Giarre, M., Zheng, Z., Brown, A. M. & Kitajewski, J. (1997) *Cell Growth Differ.* **8**, 1349–1358.
- Hashimoto, Y., Shindo-Okada, N., Tani, M., Nagamachi, Y., Takeuchi, K., Shiroishi, T., Toma, H. & Yokota, J. (1998) *J. Exp. Med.* **187**, 289–296.
- Zhang, R., Averboukh, L., Zhu, W., Zhang, H., Jo, H., Dempsey, P. J., Coffey, R. J., Pardee, A. B. & Liang, P. (1998) *Mol. Cell Biol.* **18**, 6131–6141.
- Grotendorst, G. R. (1997) *Cytokine Growth Factor Rev.* **8**, 171–179.
- Kireeva, M. L., Mo, F. E., Yang, G. P. & Lau, L. F. (1996) *Mol. Cell Biol.* **16**, 1326–1334.
- Babic, A. M., Kireeva, M. L., Kolesnikova, T. V. & Lau, L. F. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 6355–6360.
- Martinerie, C., Huff, V., Joubert, I., Badzioch, M., Saunders, G., Strong, L. & Perbal, B. (1994) *Oncogene* **9**, 2729–2732.
- Bork, P. (1993) *FEBS Lett.* **327**, 125–130.
- Kim, H. S., Nagalla, S. R., Oh, Y., Wilson, E., Roberts, C. T., Jr. & Rosenfeld, R. G. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 12981–12986.
- Joliot, V., Martinerie, C., Dambrine, G., Plassiart, G., Brisac, M., Crochet, J. & Perbal, B. (1992) *Mol. Cell Biol.* **12**, 10–21.
- Mancuso, D. J., Tuley, E. A., Westfield, L. A., Worrall, N. K., Shelton-Inloes, B. B., Sorace, J. M., Alevy, Y. G. & Sadler, J. E. (1989) *J. Biol. Chem.* **264**, 19514–19527.
- Holt, G. D., Pangburn, M. K. & Ginsburg, V. (1990) *J. Biol. Chem.* **265**, 2852–2855.
- Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., van Mourik, J. A. & Pannekoek, H. (1991) *J. Cell Biol.* **113**, 195–205.
- Martinerie, C., Viegas-Pequignot, E., Guenard, I., Dutrillaux, B., Nguyen, V. C., Bernheim, A. & Perbal, B. (1992) *Oncogene* **7**, 2529–2534.
- Takahashi, E., Hori, T., O'Connell, P., Leppert, M. & White, R. (1991) *Cytogenet. Cell. Genet.* **57**, 109–111.
- Meese, E., Meltzer, P. S., Witkowski, C. M. & Trent, J. M. (1989) *Genes Chromosomes Cancer* **1**, 88–94.
- Garte, S. J. (1993) *Crit. Rev. Oncog.* **4**, 435–449.
- Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997) *Science* **276**, 1268–1272.
- Sun, P. D. & Davies, D. R. (1995) *Annu. Rev. Biophys. Biomol. Struct.* **24**, 269–291.
- Kireeva, M. L., Lam, S. C. T. & Lau, L. F. (1998) *J. Biol. Chem.* **273**, 3090–3096.
- Frazier, K. S. & Grotendorst, G. R. (1997) *Int. J. Biochem. Cell Biol.* **29**, 153–161.
- Wernert, N. (1997) *Virchows Arch.* **430**, 433–443.
- Tanner, M. M., Tirkkonen, M., Kallioniemi, A., Collins, C., Stokke, T., Karhu, R., Kowbel, D., Shadravan, F., Hintz, M., Kuo, W. L., *et al.* (1994) *Cancer Res.* **54**, 4257–4260.
- Brinkmann, U., Gallo, M., Polymeropoulos, M. H. & Pastan, I. (1996) *Genome Res.* **6**, 187–194.
- Bischoff, J. R., Anderson, L., Zhu, Y., Mossie, K., Ng, L., Souza, B., Schryver, B., Flanagan, P., Clairvoyant, F., Ginther, C., *et al.* (1998) *EMBO J.* **17**, 3052–3065.
- Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B. & Kinzler, K. W. (1997) *Science* **275**, 1787–1790.
- Lu, L. H. & Gillett, N. (1994) *Cell Vision* **1**, 169–176.

## Review

Paul A. Haynes  
Steven P. Gygi  
Daniel Figgeys  
Ruedi Aebersold

Department of Molecular  
Biotechnology, University of  
Washington, Seattle, WA, USA

## Proteome analysis: Biological assay or data archive?

In this review we examine the current state of proteome analysis. There are three main issues discussed: why it is necessary to study proteomes; how proteomes can be analyzed with current technology; and how proteome analysis can be used to enhance biological research. We conclude that proteome analysis is an essential tool in the understanding of regulated biological systems. Current technology, while still mostly limited to the more abundant proteins, enables the use of proteome analysis both to establish databases of proteins present, and to perform biological assays involving measurement of multiple variables. We believe that the utility of proteome analysis in future biological research will continue to be enhanced by further improvements in analytical technology.

### Contents

1	Introduction .....	1862
2	Rationale for proteome analysis .....	1862
2.1	Correlation between mRNA and protein expression levels .....	1863
2.2	Proteins are dynamically modified and processed .....	1863
2.3	Proteomes are dynamic and reflect the state of a biological system .....	1863
3	Description and assessment of current proteome analysis technology .....	1863
3.1	Technical requirements of proteome technology .....	1863
3.2	2D electrophoresis – mass spectrometry: a common implementation of proteome analysis .....	1864
3.3	Protein identification by LC-MS/MS, capillary LC-MS/MS and CE-MS/MS .....	1865
3.3.1	LC-MS/MS .....	1865
3.3.2	Capillary LC-MS .....	1865
3.3.3	CE-MS/MS .....	1865
3.4	Assessment of 2-DE-MS proteome technology .....	1866
4	Utility of proteome analysis for biological research .....	1868
4.1	The proteome as a database .....	1868
4.2	The proteome as a biological assay .....	1868
5	Concluding remarks .....	1870
6	References .....	1870

### 1 Introduction

A proteome has been defined as the protein complement expressed by the genome of an organism, or, in multicellular organisms, as the protein complement expressed by a tissue or differentiated cell [1]. In the most common implementation of proteome analysis the proteins extracted from the cell or tissue analyzed are separated by high

resolution two-dimensional gel electrophoresis (2-DE), detected in the gel and identified by their amino acid sequence. The ease, sensitivity and speed with which gel-separated proteins can be identified by the use of recently developed mass spectrometric techniques have dramatically increased the interest in proteome technology. One of the most attractive features of such analyses is that complex biological systems can potentially be studied in their entirety, rather than as a multitude of individual components. This makes it far easier to uncover the many complex, and often obscure, relationships between mature gene products in cells. Large-scale proteome characterization projects have been undertaken for a number of different organisms and cell types. Microbial proteome projects currently in progress include, for example: *Saccharomyces cerevisiae* [2], *Salmonella enterica* [3], *Spiroplasma melliferum* [4], *Mycobacterium tuberculosis* [5], *Ochrobactrum anthropi* [6], *Haemophilus influenzae* [7], *Synechocystis* spp. [8], *Escherichia coli* [9], *Rhizobium leguminosarum* [10], and *Dictyostelium discoideum* [11]. Proteome projects underway for tissues of more complex organisms include those for: human bladder squamous cell carcinomas [12], human liver [13], human plasma [13], human keratinocytes [12], human fibroblasts [12], mouse kidney [12], and rat serum [14]. In this manuscript we critically assess the concept of proteome analysis and the technical feasibility of establishing complete proteome maps, and discuss ways in which proteome analysis and biological research intersect.

### 2 Rationale for proteome analysis

The dramatic growth in both the number of genome projects and the speed with which genome sequences are being determined has generated huge amounts of sequence information, for some species even complete genomic sequences ([15–17]). The description of the state of a biological system by the quantitative measurement of system components has long been a primary objective in molecular biology. With recent technical advances including the development of differential display-PCR [18], cDNA microarray and DNA chip technology [19, 20] and serial analysis of gene expression (SAGE) [21, 22], it is now feasible to establish global and quantitative mRNA expression maps of cells and tissues, in which the sequence of all the genes is known, at a speed and sensitivity which is not matched by current

Correspondence: Professor Ruedi Aebersold, Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, WA, 98195, USA (Tel: +206-685-4235; Fax: +206-685-6392; E-mail: ruedi@u.washington.edu)

Abbreviations: CID, collision-induced dissociation; MS/MS, tandem mass spectrometry; SAGE, serial analysis of gene expression

Keywords: Proteome / Two-dimensional polyacrylamide gel electrophoresis / Tandem mass spectrometry

protein analysis technology. Given the long-standing paradigm in biology that DNA synthesizes RNA which synthesizes protein, and the ability to rapidly establish global, quantitative mRNA expression maps, the questions which arise are why technically complex proteome projects should be undertaken and what specific types of information could be expected from proteome projects which cannot be obtained from genomic and transcript profiling projects. We see three main reasons for proteome analysis to become an essential component in the comprehensive analysis of biological systems. (i) Protein expression levels are not predictable from the mRNA expression levels, (ii) proteins are dynamically modified and processed in ways which are not necessarily apparent from the gene sequence, and (iii) proteomes are dynamic and reflect the state of a biological system.

## 2.1 Correlation between mRNA and protein expression levels

Interpretations of quantitative mRNA expression profiles frequently implicitly or explicitly assume that for specific genes the transcript levels are indicative of the levels of protein expression. As part of an ongoing study in our laboratory, we have determined the correlation of expression at the mRNA and protein levels for a population of selected genes in the yeast *Saccharomyces cerevisiae* growing at mid-log phase (S. P. Gygi *et al.*, submitted for publication). mRNA expression levels were calculated from published SAGE frequency tables [22]. Protein expression levels were quantified by metabolic radiolabeling of the yeast proteins, liquid scintillation counting of the protein spots separated by high resolution 2-DE and mass spectrometric identification of the protein(s) migrating to each spot. The selected 80 samples constitute a relatively homogeneous group with respect to predicted half-life and expression level of the protein products. Thus far, we have found a general trend but no strong correlation between protein and transcript levels (Fig. 1). For some genes studied equivalent mRNA transcript levels translated into protein abundances which varied by more than 50-fold. Similarly, equivalent steady-state protein expression levels were maintained by transcript levels varying by as much as 40-fold (S. P. Gygi *et al.*, submitted). These results suggest that even for a population of genes predicted to be relatively homogeneous with respect to protein half-life and gene expression, the protein levels cannot be accurately predicted from the level of the corresponding mRNA transcript.

## 2.2 Proteins are dynamically modified and processed

In the mature, biologically active form many proteins are post-translationally modified by glycosylation, phosphorylation, prenylation, acylation, ubiquitination or one or more of many other modifications [23] and many proteins are only functional if specifically associated or complexed with other molecules, including DNA, RNA, proteins and organic and inorganic cofactors. Frequently, modifications are dynamic and reversible and may alter the precise three-dimensional structure and the state of activity of a protein. Collectively, the state of modification of the proteins which constitute a biological system

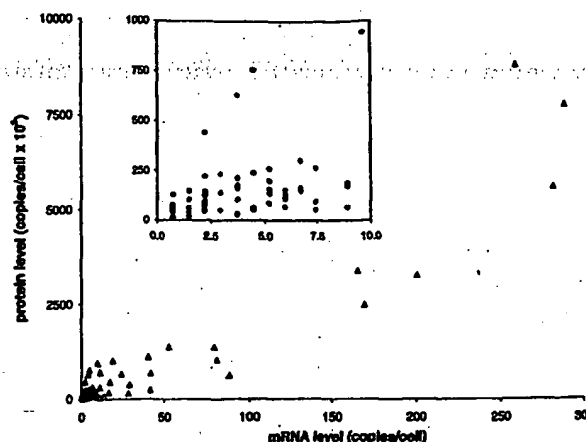


Figure 1. Correlation between mRNA and protein levels in yeast cells. For a selected population of 80 genes, protein levels were measured by  $^{35}\text{S}$ -radiolabeling and mRNA levels were calculated from published SAGE tables. Inset: expanded view of the low abundance region. For more experimental details, also see Figs. 5 and 6, (S. P. Gygi *et al.*, submitted).

are important indicators for the state of the system. The type of protein modification and the sites modified at a specific cellular state can usually not be determined from the gene sequence alone.

## 2.3 Proteomes are dynamic and reflect the state of a biological system

A single genome can give rise to many qualitatively and quantitatively different proteomes. Specific stages of the cell cycle and states of differentiation, responses to growth and nutrient conditions, temperature and stress, and pathological conditions represent cellular states which are characterized by significantly different proteomes. The proteome, in principle, also reflects events that are under translational and post-translational control. It is therefore expected that proteomics will be able to provide the most precise and detailed molecular description of the state of a cell or tissue, provided that the external conditions defining the state are carefully determined. In answer to the question of whether the study of proteomes is necessary for the analysis of biomolecular systems, it is evident that the analysis of mature protein products in cells is essential as there are numerous levels of control of protein synthesis, degradation, processing and modification, which are only apparent by direct protein analysis.

## 3 Description and assessment of current proteome analysis technology

### 3.1 Technical requirements of proteome technology

In biological systems the level of expression as well as the states of modification, processing and macro-molecular association of proteins are controlled and modulated depending on the state of the system. Comprehensive analysis of the identity, quantity and state of modification of proteins therefore requires the detection and

quantitation of the proteins which constitute the system, and analysis of differentially processed forms. There are a number of inherent difficulties in protein analysis which complicate these tasks. First, proteins cannot be amplified. It is possible to produce large amounts of a particular protein by over-expression in specific cell systems. However, since many proteins are dynamically post-translationally modified, they cannot be easily amplified in the form in which they finally function in the biological system. It is frequently difficult to purify from the native source sufficient amounts of a protein for analysis. From a technological point of view this translates into the need for high sensitivity analytical techniques. Second, many proteins are modified and processed post-translationally. Therefore, in addition to the protein identity, the structural basis for differentially modified isoforms also needs to be determined. The distribution of a constant amount of protein over several differentially modified isoforms further reduces the amount of each species available for analysis. The complexity and dynamics of post-translational protein editing thus significantly complicates proteome studies. Third, proteins vary dramatically with respect to their solubility in commonly used solvents. There are few, if any, solvent conditions in which all proteins are soluble and which are also compatible with protein analysis. This makes the development of protein purification methods particularly difficult since both protein purification and solubility have to be achieved under the same conditions. Detergents, in particular sodium dodecyl sulfate (SDS), are frequently added to aqueous solvents to maintain protein solubility. The compatibility with SDS is a big advantage of SDS polyacrylamide gel electrophoresis (SDS-PAGE) over other protein separation techniques. Thus, SDS-PAGE and two-dimensional gel electrophoresis, which also uses SDS and other detergents, are the most general and preferred methods for the purification of small amounts of proteins, provided that activity does not necessarily need to be maintained. Lastly, the number of proteins in a given cell system is typically in the thousands. Any attempt to identify and categorize all of these must use methods which are as rapid as possible to allow completion of the project within a reasonable time frame. Therefore, a successful, general proteomics technology requires high sensitivity, high throughput, the ability to differentiate differentially modified proteins, and the ability to quantitatively display and analyze all the proteins present in a sample.

### 3.2 2-D electrophoresis – mass spectrometry: a common implementation of proteome analysis

The most common currently used implementation of proteome analysis technology is based on the separation of proteins by two-dimensional (IEF/SDS-PAGE) gel electrophoresis and their subsequent identification and analysis by mass spectrometry (MS) or tandem mass spectrometry (MS/MS). In 2-DE, proteins are first separated by isoelectric focusing (IEF) and then by SDS-PAGE, in the second, perpendicular dimension. Separated proteins are visualized at high sensitivity by staining or autoradiography, producing two-dimensional arrays of proteins. 2-DE gels are, at present, the most commonly used means of global display of proteins in complex

samples. The separation of thousands of proteins has been achieved in a single gel [24, 25] and differentially modified proteins are frequently separated. Due to the compatibility of 2-DE with high concentrations of detergents, protein denaturants and other additives promoting protein solubility, the technique is widely used.

The second step of this type of proteome analysis is the identification and analysis of separated proteins. Individual proteins from polyacrylamide gels have traditionally been identified using *N*-terminal sequencing [26, 27], internal peptide sequencing [28, 29], immunoblotting or comigration with known proteins [30]. The recent dramatic growth of large-scale genomic and expressed sequence tag (EST) sequence databases has resulted in a fundamental change in the way proteins are identified by their amino acid sequence. Rather than by the traditional methods described above, protein sequences are now frequently determined by correlating mass spectral or tandem mass spectral data of peptides derived from proteins, with the information contained in sequence databases [31–33].

There are a number of alternative approaches to proteome analysis currently under development. There is considerable interest in developing a proteome analysis strategy which bypasses 2-DE altogether, because it is considered a relatively slow and tedious process, and because of perceived difficulties in extracting proteins from the gel matrix for analysis. However, 2-DE as a starting point for proteome analysis has many advantages compared to other techniques available today. The most significant strengths of the 2-DE-MS approach include the relatively uniform behavior of proteins in gels, the ability to quantify spots and the high resolution and simultaneous display of hundreds to thousands of proteins within a reasonable time frame.

A schematic diagram of a typical procedure of the identification of gel-separated proteins is shown in Fig. 2. Protein spots detected in the gel are enzymatically or chemically fragmented and the peptide fragments are isolated for analysis, as already indicated, most frequently by MS or MS/MS. There are numerous protocols for the generation of peptide fragments from gel-separated proteins. They can be grouped into two categories, digestion in the gel slice [28, 34] or digestion after electrotransfer out of the gel onto a suitable membrane ([29, 35–37] and reviewed in [38]). In most instances either technique is applicable and yields good results. The analysis of MS or MS/MS data is an important step in the whole process because MS instruments can generate an enormous amount of information which cannot easily be managed manually. Recently, a number of groups have developed software systems dedicated to the use of peptide MS and MS/MS spectra for the identification of proteins. Proteins are identified by correlating the information contained in the MS spectra of protein digests or MS/MS spectra of individual peptides with data contained in DNA or protein sequence databases.

The systems we are currently using in our laboratory are based on the separation of the peptides contained in protein digests by narrow bore or capillary liquid chromatog-



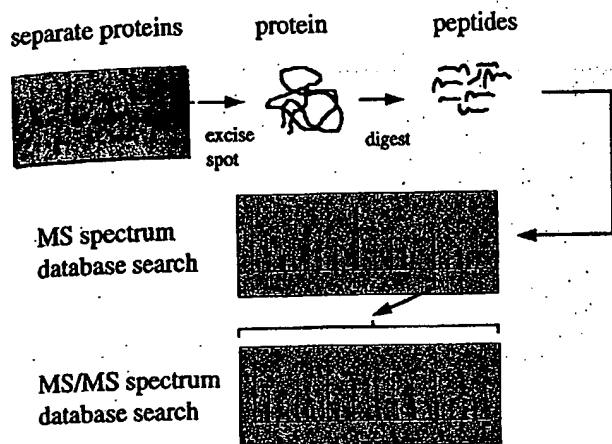


Figure 2. Schematic diagram of a procedure for identification of gel-separated proteins. Peptides can either be separated by a technique such as LC or CE, or infused as a mixture and sorted in the MS. Database searching can either be performed on peptide masses from an MS spectrum, peptide fragment masses from CID spectra of peptides, or a combination of both.

raphy [39, 40] or capillary electrophoresis [41], the analysis of the separated peptides by electrospray ionization (ESI) MS/MS, and the correlation of the generated peptide spectra with sequence databases using the SEQUEST program developed at the University of Washington [32, 33]. The system automatically performs the following operations: a particular peptide ion characterized by its mass-to-charge ratio is selected in the MS out of all the peptide ions present in the system at a particular time; the selected peptide ion is collided in a collision cell with argon (collision-induced dissociation, CID) and the masses of the resulting fragment ions are determined in the second sector of the tandem MS; this experimentally determined CID spectrum is then correlated with the CID spectra predicted from all the peptides in a sequence database which have essentially the same mass as the peptide selected for CID; this correlation matches the isolated peptide with a sequence segment in a database and thus identifies the protein from which the peptide was derived. There are a number of alternative programs which use peptide CID spectra for protein identification, but we use the SEQUEST system because it is currently the most highly automated program and has proven to be successful, versatile and robust.

### 3.3 Protein identification by LC-MS/MS, capillary LC-MS/MS and CE-MS/MS

It has been demonstrated repeatedly that MS has a very high intrinsic sensitivity. For the routine analysis of gel-separated proteins at high sensitivity, the most significant challenge is the handling of small amounts of sample. The crux of the problem is the extraction and transfer of peptide mixtures generated by the digestion of low nanogram amounts of protein, from gels into the MS/MS system without significant loss of sample or introduction of unwanted contaminants. We employ three different systems for introducing gel-purified samples into an MS, depending on the level of sensitivity

required. As an approximate guideline, for samples containing tens of picomoles of peptides, LC-MS/MS is most appropriate; for samples containing low picomole amounts to high femtomole amounts we use capillary LC-MS/MS; and for samples containing femtomoles or less, CE-MS/MS is the method of choice.

#### 3.3.1 LC-MS/MS

The coupling of an MS to an HPLC system using a 0.5 mm diameter or bigger reverse phase (RP) column has been described in detail [42]. This system has several advantages if a large number of samples are to be analyzed and all are available in sufficient quantity. The LC-MS and database searching program can be run in a fully automated mode using an autosampler, thus maximizing sample throughput and minimizing the need for operator interference. The relatively large column is tolerant of high levels of impurities from either gel preparation or sample matrix. Lastly, if configured with a flow-splitter and micro-sprayer [40], analyses can be performed on a small fraction of the sample (less than 5%) while the remainder of the sample is recovered in very pure solvents. This latter feature is particularly useful when an orthogonal technique is also used to analyze peptide fractions, such as scintillation of an introduced radiolabel, and this data can be correlated with peptides identified by CID spectra.

#### 3.3.2 Capillary LC-MS

An increase of sensitivity of approximately tenfold can be achieved by using a capillary LC system with a 100  $\mu$ m ID column rather than a 0.5 mm ID column as referred to above. Since very low flow rates are required for such columns, most reports have used a precolumn flow splitting system for producing solvent gradients. We have recently described the design and construction of a novel gradient mixing system which enables the formation of reproducible gradients at very low flow rates (low nL/min) without the need for flow splitting (A. Ducret *et al.*, submitted for publication). Using this capillary LC-MS/MS system we were able to identify gel-separated proteins if low picomole to high femtomole amounts were loaded onto the gel [40]. This system is as yet not automated and, like all capillary LC systems, is prone to blockage of the columns by microparticulates when analyzing gel-separated proteins.

#### 3.3.3 CE-MS/MS

The highest level of sensitivity for analyzing gel-separated proteins can be achieved by using capillary electrophoresis – mass spectrometry (CE-MS). We have described in the past a solid-phase extraction capillary electrophoresis (SPE-CE) system which was used with triple quadrupole and ion trap ESI-MS/MS systems for the identification of proteins at the low femtomole to sub-femtomole sensitivity level [43, 44]. While this system is highly sensitive, its operation is labor-intensive and its operation has not been automated. In order to devise an analytical system with both the sensitivity of a CE and the level of automation of LC, we have constructed

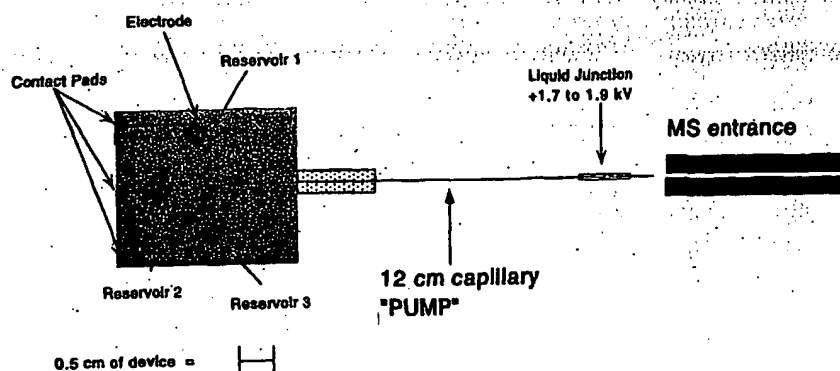


Figure 3. Schematic illustration of a microfabricated analytical system for CE, consisting of a micromachined device, coated capillary electroosmotic pump, and microelectrospray interface. The dimensions of the channels and reservoir are as indicated in the text. The channels on the device were graphically enhanced to make them more visible. Reproduced from [45], with permission.

microfabricated devices for the introduction of samples into ESI-MS for high-sensitivity peptide analysis.

The basic device is a piece of glass into which channels of 10–30  $\mu\text{m}$  in depth and 50–70  $\mu\text{m}$  in diameter are etched by using photolithography/etching techniques similar to the ones used in the semiconductor industry. (A simple device is shown in Fig. 3). The channels are connected to an external high voltage power supply [45]. Samples are manipulated on the device and off the device to the MS by applying different potentials to the reservoirs. This creates a solvent flow by electroosmotic pumping which can be redirected by changing the position of the electrode. Therefore, without the need for valves or gates and without any external pumping, the flow can be redirected by simply switching the position of the electrodes on the device. The direction and rate of the flow can be modulated by the size and the polarity of the electric field applied and also by the charge state of the surface.

The type of data generated by the system is illustrated in Fig. 4, which shows the mass spectrum of a peptide sample representing the tryptic digest of carbonic anhydrase at 290 fmol/ $\mu\text{L}$ . Each numbered peak indicates a peptide successfully identified as being derived from carbonic an-

hydrase. Some of the unassigned signals may be chemical or peptide contaminants. The MS is programmed to automatically select each peak and subject the peptide to CID. The resulting CID spectra are then used to identify the protein by correlation with sequence databases. Therefore, this system allows us to concurrently apply a number of protein digests onto the device, to sequentially mobilize the samples, to automatically generate CID spectra of selected peptide ions and to search sequence databases for protein identification. These steps are performed automatically without the need for user input and proteins can be identified at very low femtomole level sensitivity at a rate of approximately one protein per 15 min.

#### 3.4 Assessment of 2-DE-MS proteome technology

Using a combination of the analytical techniques described above we have identified the 80 protein spots indicated in Fig. 5. The protein pattern was generated by separating a total of 40 microgram of protein contained in a total cell lysate of the yeast strain YPH499 by high resolution 2-DE and silver staining of the separated proteins. To estimate how far this type of proteome analysis can penetrate towards the identification of low abundance proteins, we have calculated the codon bias of the genes encoding the respective proteins. Codon bias is a

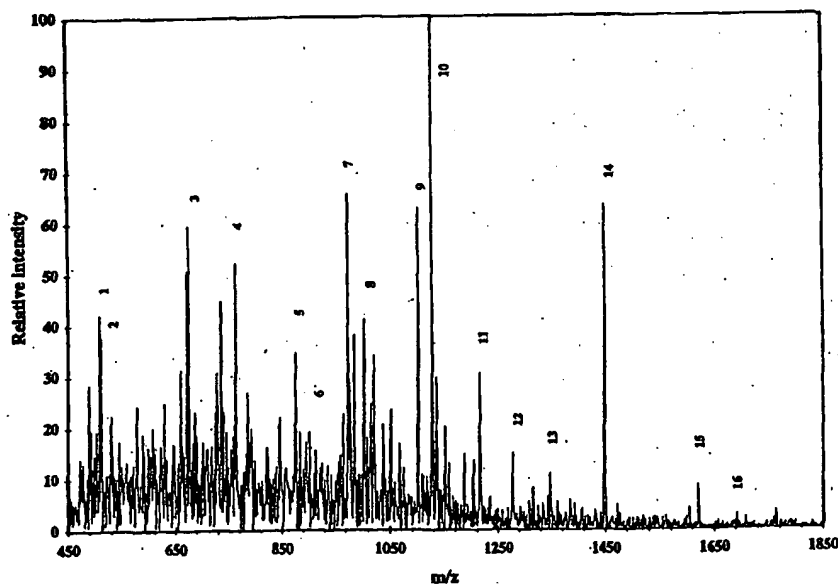


Figure 4. MS spectrum of a tryptic digest of carbonic anhydrase using the microfabricated system shown in Fig. 3. 290 fmol/ $\mu\text{L}$  of carbonic anhydrase tryptic digest was infused into a Finnigan LCQ ion trap MS. Each peak was selected for CID, and those which were identified as containing peptides derived from carbonic anhydrase are numbered. Reproduced from [45], with permission.



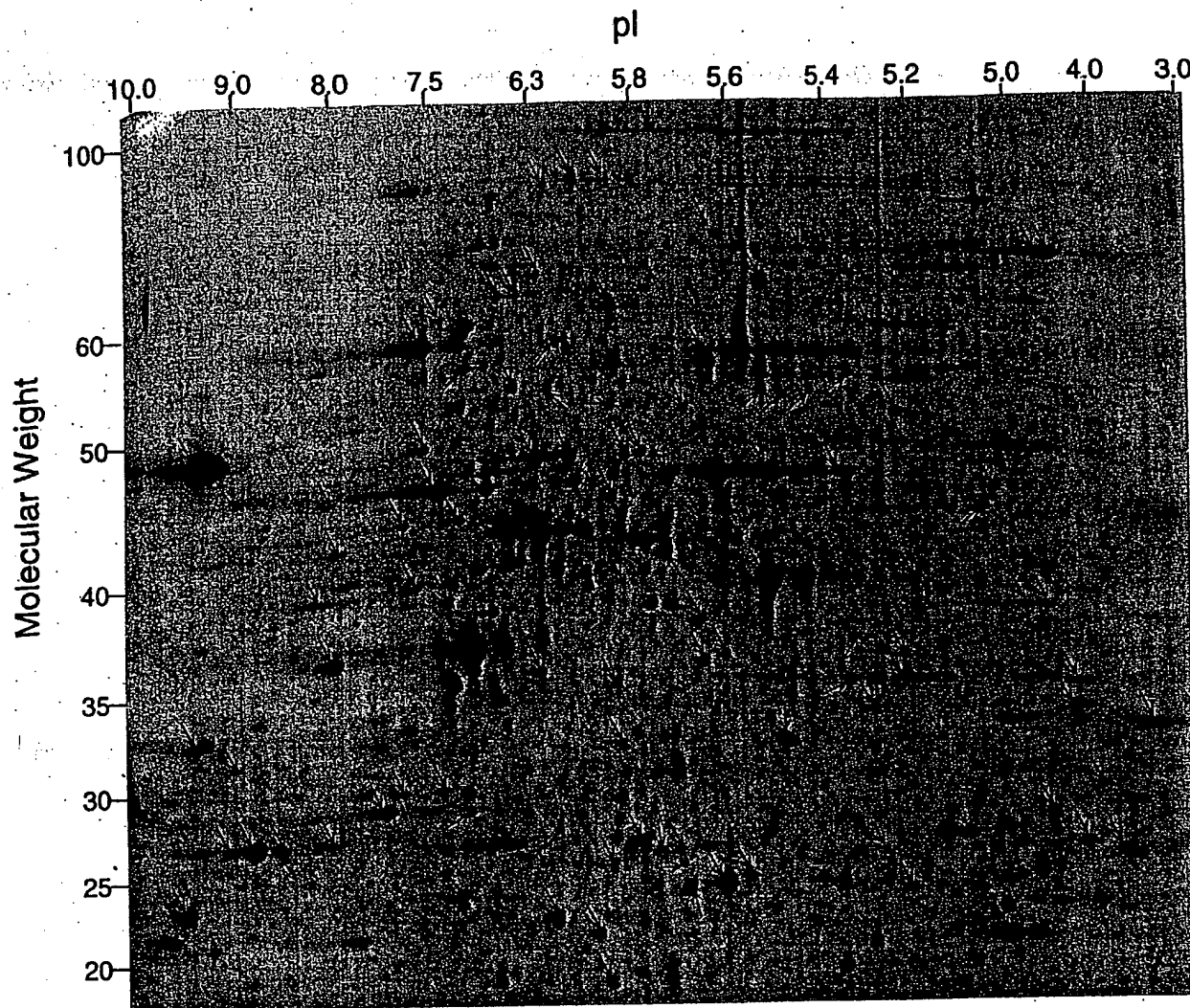


Figure 5. 2-DE separation of a lysate of yeast cells, with identified proteins highlighted. The first dimension of separation was an IPG from pH 3–10, and the second dimension was a 10%T SDS-PAGE gel. Proteins were visualized by silver staining. Further details of experimental procedures are included in S. P. Gygi *et al.* (submitted).

calculated measure of the degree of redundancy of triplet DNA codons used to produce each amino acid in a particular gene sequence. It has been shown to be a useful indicator of the level of the protein product of a particular gene sequence present in a cell [46]. The general rule which applies is that the higher the value of the codon bias calculated for a gene, the more abundant the protein product of that gene becomes. The calculated codon bias values corresponding to the proteins identified in Fig. 5 are shown in Fig. 6b. Nearly all of the proteins identified (> 95%) have codon bias values of > 0.2, indicating they are highly abundant in cells. In contrast, codon bias values calculated for the entire yeast genome (Fig. 6a) show that the majority of proteins present in the proteome have a codon bias of < 0.2 and are thus of low abundance.

This finding is of considerable importance in our assessment of the current status of proteome analysis technology. It is clear that even using highly sensitive analytical techniques, we are only able to visualize and identify the

more abundant proteins. Since many important regulatory proteins are present only at low abundance, these would not be amenable to analysis using such techniques. This situation would be exacerbated in the analysis of proteomes containing many more proteins than the approximately 6000 gene products present in yeast cells [16]. In the analysis of, for example, the proteome of any human cells, there are potentially 50 000–100 000 gene products [47]. Inherent limitations on the amount of protein that can be loaded on 2-DE, and the number of components that can be resolved, indicate that only the most highly abundant fraction of the many gene products could be successfully analyzed. One approach that has been employed to circumvent these limitations is the use of very narrow range immobilized pH gradient strips for the first-dimension separation of 2-DE [48]. Since only those proteins which focus within the narrow range will enter the second dimension of separation, a much higher sample loading within the desired range is possible. This, in turn, can lead to the visualization and identification of less abundant proteins.

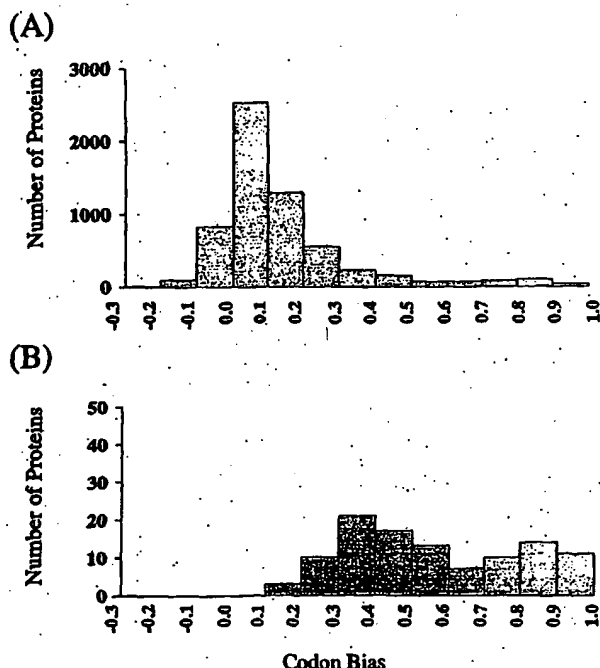


Figure 6. Calculated codon bias values for yeast proteins. (A) Distribution of calculated values for the entire yeast proteome. (B) Distribution of calculated values for the subset of 80 identified proteins also shown in Figs. 1 and 5. Further details of experimental procedures are included in S. P. Gygi *et al.* (submitted).

#### 4 Utility of proteome analysis for biological research

For the success of proteomics as a mainstream approach to the analysis of biological systems it is essential to define how proteome analysis and biological research projects intersect. Without a clear plan for the implementation of proteome-type approaches into biological research projects the full impact of the technology can not be realized. The literature indicates that proteome analysis is used both as a database/data archive, and as a biological assay or biological research tool.

##### 4.1 The proteome as a database

The use of proteomics as a database or data archive essentially entails an attempt to identify all the proteins in a cell or species and to annotate each protein with the known biological information that is relevant for each protein. The level of annotation can, of course, be extensive. The most common implementation of this idea is the separation of proteins by high resolution 2-DE, the identification of each detected protein spot and the annotation of the protein spots in a 2-DE gel database format. This approach is complicated by the fact that it is difficult to precisely define a proteome and to decide which proteome should be represented in the database. In contrast to the genome of a species, which is essentially static, the proteome is highly dynamic. Processes such as differentiation, cell activation and disease can all significantly change the proteome of a species. This is illustrated in Fig. 7. The figure shows two high-resolu-

tion 2-DE maps of proteins isolated from rat serum. Fig. 7A is from the serum of normal rats, while Fig. 7B is from the serum of rats in acute-phase serum after prior treatment with an inflammation-causing agent [49]. It is obvious that the protein patterns are significantly different in several areas, raising the question of exactly which proteome is being described.

Therefore, a comprehensive proteome database of a species or cell type needs to contain all of the parameters which describe the state and the type of the cells from which the proteins were extracted as well as the software tools to search the database with queries which reflect the dynamics of biological systems. A comprehensive proteome database should be capable of quantitatively describing the fate of each protein if specific systems and pathways are activated in the cell. Specifically, the quantity, the degree of modification, the subcellular location and the nature of molecules specifically interacting with a protein as well as the rate of change of these variables should be described. Using these admittedly stringent criteria, there is currently no complete proteome database. A number of such databases are, however, in the process of being constructed. The most advanced among them, in our opinion, are the yeast protein database YPD [50] (accessible at <http://www.ypd.com>) and the human 2D-PAGE databases of the Danish Centre for Human Genome Research [12] (accessible at <http://biobase.dk/cgi-bin/celisis>). While neither can be considered complete as not all of the potential gene products are identified, both contain extensive annotation of supplemental information for many of the spots which are positively identified in reference samples.

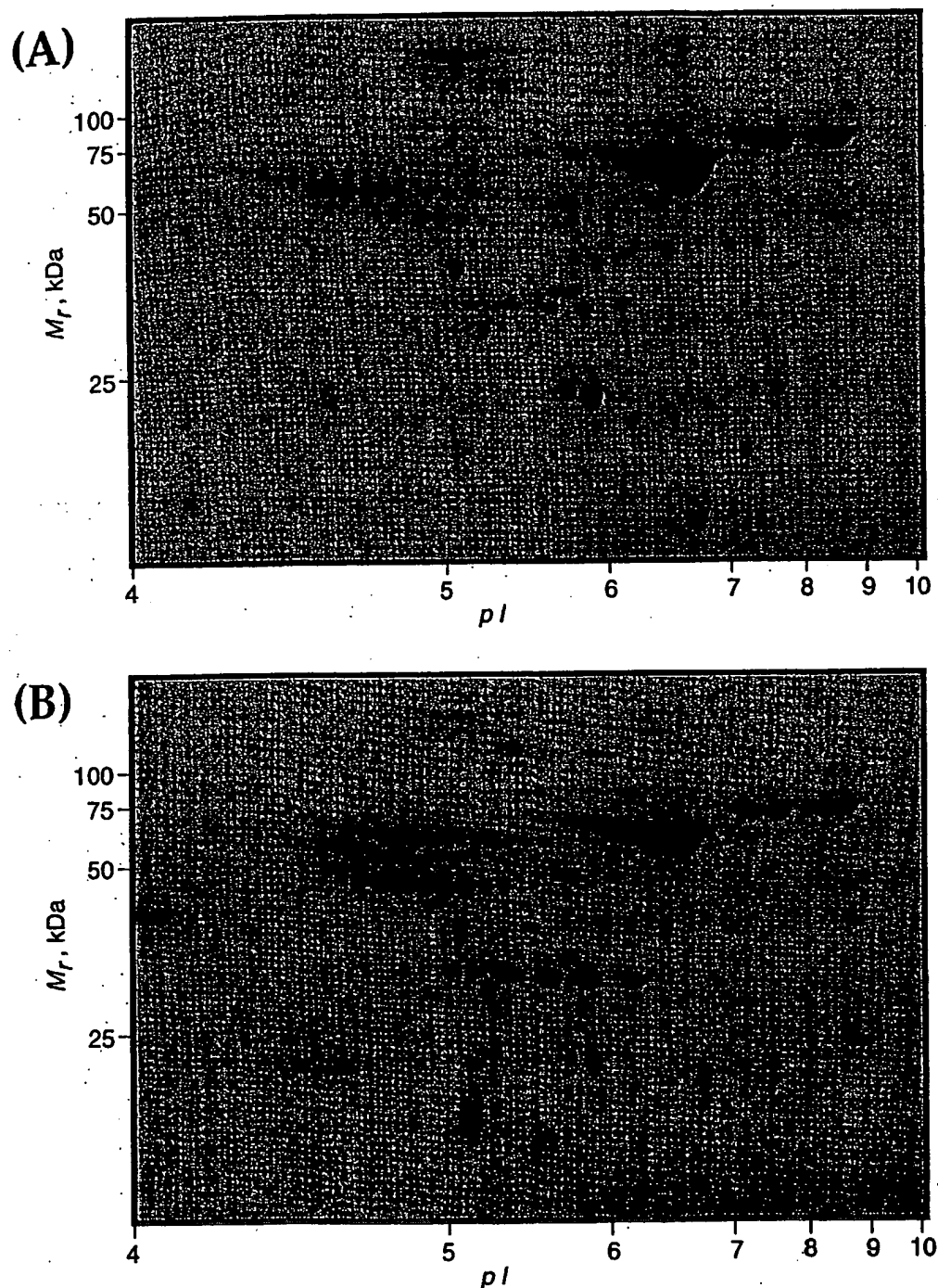
##### 4.2 The proteome as a biological assay

The use of proteome analysis as a biological assay or research tool represents an alternative approach to integrating biology with proteomics. To investigate the state of a system, samples are subjected to a specific process that allows the quantitative or qualitative measurement of some of the variables which describe the system. In typical biochemical assays one variable (e.g., enzyme activity) of a single component (e.g., a particular enzyme) is measured. Using proteomics as an assay, multiple variables (e.g., expression level, rate of synthesis, phosphorylation state, etc.) are measured concurrently on many (ideally all) of the proteins in a sample. The use of proteomics as an assay is a less far-reaching proposition than the construction of a comprehensive proteome database. It does, however, represent a pragmatic approach which can be adapted to investigate specific systems and pathways, as long as the interpretation of the results takes into account that with current technology not all of the variables which describe the system can be observed (see Section 3.4).

A common implementation of proteome analysis as a biological assay is when a 2-DE protein pattern generated from the analysis of an experimental sample is compared to an array of reference patterns representing different states of the system under investigation. The state of the experimental system at the time the sample was generated is therefore determined by the quantita-

tive comparative analysis of hundreds to a few thousand proteins. Comparative analysis of the 2-DE patterns furthermore highlights quantitative and qualitative differences in the protein profiles which correlate with the state of the system. For this type of analysis it is not essential that all the proteins are identified or even visu-

alized, although the results become more informative as more proteins are compared. It is obvious, however, that the possibility to identify any protein deemed characteristic for a particular state dramatically enhances this approach by opening up new avenues for experimentation.



**Figure 7.** High resolution 2-DE map of proteins isolated from rat serum with or without prior exposure to an inflammation-causing agent. (A) normal rat serum, (B) acute-phase serum from rats which had previously been exposed to an inflammation-causing agent. The first dimension of separation is an IPG from pH 4–10, and the second dimension is a 7.5–17.5%T gradient SDS-PAGE gel. Proteins were visualized by staining with amido black. Further details of experimental procedures are included in [14, 49].

Proteome analysis as a biological assay has been successfully used in the field of toxicology, to characterize disease states or to study differential activation of cells. The approach is limited, of course, by the fact that only the visible protein spots are included in the assay, and it is well known that a substantial but far from complete fraction of cellular proteins are detected if a total cell lysate is separated by 2-DE. Proteins may not be detected in 2-DE gels because they are not abundant enough to be visualized by the detection method used, because they do not migrate within the boundaries (size, pI) resolved by the gel, because they are not soluble under the conditions used, or for other reasons.

A different way to use proteome analysis as a biological assay to define the state of a biological system is to take advantage of the wealth of information contained in 2-DE protein patterns. 2-DE is referred to as two-dimensional because of the electrophoretic mobility and the isoelectric points which define the position of each protein in a 2-DE pattern. In addition to the two dimensions used to generate the protein patterns, a number of additional data dimensions are contained in the protein patterns. Some of these dimensions such as protein expression level, phosphorylation state, subcellular location, association with other proteins, rate of synthesis or degradation indicate the activity state of a protein or a biological system. Comparative analysis of 2-DE protein patterns representing different states is therefore ideally suited for the detection, identification and analysis of suitable markers. Once again it must be emphasized that in this type of experiment only a fraction of the cellular proteins is analyzed. Since many regulatory proteins are of low abundance, this limitation is a concern, particularly in cases in which regulatory pathways are being investigated.

## 5 Concluding remarks

In this report we have addressed three main issues related to proteome analysis. First, we have discussed the rationale for studying proteomes. Second, we have assessed the technical feasibility of analyzing proteomes and described current proteome technology, and third, we have analyzed the utility of proteome analysis for biological research. It is apparent that proteome analysis is an essential tool in the analysis of biological systems. The multi-level control of protein synthesis and degradation in cells means that only the direct analysis of mature protein products can reveal their correct identities, their relevant state of modification and/or association and their amounts. Recently developed methods have enabled the identification of proteins at ever-increasing sensitivity levels and at a high level of automation of the analytical processes. A number of technical challenges, however, remain. While it is currently possible to identify essentially any protein spots that can be visualized by common staining methods, it is apparent that without prior enrichment only a relatively small and highly selected population of long-lived, highly expressed proteins is observed. There are many more proteins in a given cell which are not visualized by such methods. Frequently it is the low abundance proteins that execute key regulatory functions.

We have outlined the two principal ways proteome analysis is currently being used to intersect with biological research projects: the proteome as a database or data archive and proteome analysis as a biological assay. Both approaches have in common that at present they are conceptually and technically limited. Current proteome databases typically are limited to one cell type and one state of a cell and therefore do not account for the dynamics of biological systems. The use of proteome analysis as a biological assay can provide a wealth of information, but it is limited to the proteins detected and is therefore not truly proteome-wide. These limitations in proteomics are to a large extent a reflection of the fact that proteins in their fully processed form cannot easily be amplified and are therefore difficult to isolate in amounts sufficient for analysis or experimentation. The fact that to date no complete proteome has been described further attests to these difficulties. With continued rapid progress in protein analysis technology, however, we anticipate that the goal of complete proteome analysis will eventually become attainable.

*We would like to acknowledge the funding for our work from the National Science Foundation Science and Technology Center for Molecular Biotechnology and from the NIH. We thank Yvan Rochon and Bob Franza for providing the yeast gel shown and Elisabetta Gianazza for providing the rat serum gels shown.*

Received April 21, 1998

## 6 References

- [1] Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., Hochstrasser, D. F., *Bio/Technology* 1996, 14, 61-65.
- [2] Hodges, P. E., Payne, W. E., Garrels, J. I., *Nucleic Acids Res.* 1998, 26, 68-72.
- [3] O'Connor, C. D., Farris, M., Fowler, R., Qi, S. Y., *Electrophoresis* 1997, 18, 1483-1490.
- [4] Cordwell, S. J., Basseal, D. J., Humphery-Smith, I., *Electrophoresis* 1997, 18, 1335-1346.
- [5] Urquhart, B. L., Atsalos, T. E., Roach, D., Basseal, D. J., Bjellqvist, B., Britton, W. L., Humphery-Smith, I., *Electrophoresis* 1997, 18, 1384-1392.
- [6] Wasinger, V. C., Bjellqvist, B., Humphery-Smith, I., *Electrophoresis* 1997, 18, 1373-1383.
- [7] Link, A. J., Hays, L. G., Carmack, E. B., Yates III, J. R., *Electrophoresis* 1997, 18, 1314-1334.
- [8] Sazuka, T., Ohara, O., *Electrophoresis* 1997, 18, 1252-1258.
- [9] VanBogelen, R. A., Abshire, K. Z., Moldover, B., Olson, E. R., Neidhardt, F. C., *Electrophoresis* 1997, 18, 1243-1251.
- [10] Guerreiro, N., Redmond, J. W., Rolfe, B. G., Djordjevic, M. A., *Mol. Plant Microbe Interact.* 1997, 10, 506-516.
- [11] Yan, J. X., Tonella, L., Sanchez, J.-C., Wilkins, M. R., Packer, N. H., Gooley, A. A., Hochstrasser, D. F., Williams, K. L., *Electrophoresis* 1997, 18, 491-497.
- [12] Celis, J., Gromov, P., Ostergaard M., Madsen, P., Honoré, B., Deigaard, K., Olsen, E., Vorum, H., Kristensen, D. B., Gromova, I., Haunso, A., Van Damme, J., Puype, M., Vandekerckhove, J., Rasmussen, H. H., *FEBS Lett.* 1996, 398, 129-134.
- [13] Appel, R. D., Sanchez, J.-C., Bairoch, A., Golaz, O., Miu, M., Vargas, J. R., Hochstrasser, D. F., *Electrophoresis* 1993, 14, 1232-1238.
- [14] Haynes, P., Miller, I., Aebersold, R., Gemeiner, M., Eberini, I., Lovati, R. M., Manzoni, C., Vignati, M., Gianazza, E., *Electrophoresis* 1998, 19, 1484-1492.

- [15] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, N. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, C. O., Venter, J. C., *Science* 1996, 269, 496-512.
- [16] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S. G., *Science* 1996, 274, 546.
- [17] Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., Peterson, J., Kerlavage, A. R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M. D., Gocayne, J., Weidman, J., Utterback, T., Watthey, T., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M. D., Horst, K., Roberts, K., Hatch, B., Smith, H. O., Venter, J. C., *Nature* 1997, 390, 580-586.
- [18] Liang, P., Pardee, A. B., *Science* 1992, 257, 967-971.
- [19] Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., Davis, R. W., *Proc. Natl. Acad. Sci. USA* 1997, 94, 13057-13062.
- [20] Shalon, D., Smith, S. J., Brown, P. O., *Genome Res.* 1996, 6, 639-645.
- [21] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., *Science* 1995, 270, 484-487.
- [22] Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., Kinzler, K. W., *Cell* 1997, 88, 243-251.
- [23] Krishna, R. G., Wold, P., *Adv. Enzymol.* 1993, 67, 265-298.
- [24] Örg, A., Postel, W., Gunther, S., *Electrophoresis* 1988, 9, 531-546.
- [25] Klose, J., Kobalz, U., *Electrophoresis* 1995, 16, 1034-1059.
- [26] Matsudaira, P., *J. Biol. Chem.* 1987, 262, 10035-10038.
- [27] Aebersold, R. H., Teplow, D. B., Hood, L. E., Kent, S. B., *J. Biol. Chem.* 1986, 261, 4229-4238.
- [28] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., *Anal. Biochem.* 1992, 203, 173-179.
- [29] Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., Kent, S. B., *Proc. Natl. Acad. Sci. USA* 1987, 84, 6970-6974.
- [30] Honoré, B., Leffers, H., Madsen, P., Celis, J. E., *Eur. J. Biochem.* 1993, 218, 421-430.
- [31] Mann, M., Wilm, M., *Anal. Chem.* 1994, 66, 4390-4399.
- [32] Eng, J., McCormack, A. L., Yates III, J. R., *J. Amer. Mass Spectrom.* 1994, 5, 976-989.
- [33] Yates III, J. R., Eng, J. K., McCormack, A. L., Schieltz, D., *Anal. Chem.* 1995, 67, 1426-1436.
- [34] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, 68, 850-858.
- [35] Hess, D., Covey, T. C., Winz, R., Brownsey, R. W., Aebersold, R., *Protein Sci.* 1993, 2, 1342-1351.
- [36] van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Biochem.* 1997, 247, 310-318.
- [37] Lui, M., Tempst, P., Erdjument-Bromage, H., *Anal. Biochem.* 1996, 241, 156-166.
- [38] Patterson, S. D., Aebersold, R. A., *Electrophoresis* 1995, 16, 1791-1814.
- [39] Ducret, A., Foyn, Brunn, C., Bures, E. J., Marhaug, G., Husby, G. R. A., *Electrophoresis* 1996, 17, 866-876.
- [40] Haynes, P. A., Fripp, N., Aebersold, R., *Electrophoresis* 1998, 19, 939-945.
- [41] Figeys, D., Van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Chem.* 1996, 68, 1822-1828.
- [42] Ducret, A., Van Oostveen, I., Eng, J. K., Yates III, J. R., Aebersold, R., *Protein Sci.* 1997, 7, 706-719.
- [43] Figeys, D., Ducret, A., Yates III, J. R., Aebersold, R., *Nature Biotech.* 1996, 14, 1579-1583.
- [44] Figeys, D., Aebersold, R., *Electrophoresis* 1997, 18, 360-368.
- [45] Figeys, D., Ning, Y., Aebersold, R., *Anal. Chem.* 1997, 69, 3153-3160.
- [46] Garrels, J. I., McLaughlin, C. S., Warner, J. R., Fletcher, B., Latter, G. I., Kobayashi, R., Schwender, B., Volpe, T., Anderson, D. S., Mesquita-Fuentes, R., Payne, W. E., *Electrophoresis* 1997, 18, 1347-1360.
- [47] Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., Edwards, C., Fan, J.-B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hul, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matise, T. C., McKusick, K. B., Morissette, J., Mungall, A., Muelet, D., Nusbaum, H. C., Page, D. C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, X., She, X., Silva, J., Slonim, D. K., Soderlund, C., Sun, W.-L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M. D., Auffray, C., Walter, N. A. R., Brandon, R., Dehejia, A., Goodfellow, P. N., Houlgate, R., Hudson, J. R., Jr., Ide, S. E., Iorio, K. R., Lee, W. Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M. H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J. C., Sikela, J. M., Beckmann, J. S., Weissenbach, J., Myers, R. M., Cox, D. R., James, M. R., Bentley, D., et al. *Science* 1996, 274, 540-546.
- [48] Sanchez, J.-C., Rouge, V., Pisteur, M., Ravier, F., Tonella, L., Moosmayer, M., Wilkins, M. R., Hochstrasser, D. F., *Electrophoresis* 1997, 18, 324-327.
- [49] Miller, I., Haynes, P., Gemeiner, M., Aebersold, R., Manzoni, C., Lovati, M. R., Vignati, M., Eberini, I., Gianazza, E., *Electrophoresis* 1998, 19, 1493-1500.
- [50] Garrels, J. I., *Nucleic Acids Res.* 1996, 24, 46-49.

## Analysis of Genomic and Proteomic Data Using Advanced Literature Mining

Yanhui Hu, Lisa M. Hines, Haifeng Weng, Dongmei Zuo, Miguel Rivera,  
Andrea Richardson, and Joshua LaBaer\*

*Institute of Proteomics, Harvard Medical School-BCMP, 240 Longwood Avenue, Boston, Massachusetts 02115*

Received March 13, 2003

High-throughput technologies, such as proteomic screening and DNA micro-arrays, produce vast amounts of data requiring comprehensive analytical methods to decipher the biologically relevant results. One approach would be to manually search the biomedical literature; however, this would be an arduous task. We developed an automated literature-mining tool, termed MedGene, which comprehensively summarizes and estimates the relative strengths of all human gene-disease relationships in Medline. Using MedGene, we analyzed a novel micro-array expression dataset comparing breast cancer and normal breast tissue in the context of existing knowledge. We found no correlation between the strength of the literature association and the magnitude of the difference in expression level when considering changes as high as 5-fold; however, a significant correlation was observed ( $r = 0.41$ ;  $p = 0.05$ ) among genes showing an expression difference of 10-fold or more. Interestingly, this only held true for estrogen receptor (ER) positive tumors, not ER negative. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative tumors worthy of further examination.

**Keywords:** bioinformatics • micro-array • text mining • gene-disease association • breast cancer

### Introduction

At its current pace, the accumulation of biomedical literature outpaces the ability of most researchers and clinicians to stay abreast of their own immediate fields, let alone cover a broader range of topics. For example, to follow a single disease, e.g., breast cancer, a researcher would have had to scan 130 different journals and read 27 papers per day in 1999.<sup>1</sup> This problem is accentuated with high-throughput technologies such as DNA micro-arrays and proteomics, which require the analysis of large datasets involving thousands of genes, many of which are unfamiliar to a particular researcher. In any microarray experiment, thousands of genes may demonstrate statistically significant expression changes, but only a fraction of these may be relevant to the study. The ability to interpret these datasets would be enhanced if they could be compared to a comprehensive summary of what is known about all genes. Thus, there is a need to summarize existing knowledge in a format that allows for the rapid analysis of associations between genes and diseases or other specific biological concepts.

One solution to this problem is to compile structured digital resources, such as the Breast Cancer Gene Database<sup>1</sup> and the Tumor Gene Database.<sup>2</sup> However, as these resources are hand-curated, the labor-intensive review process becomes a rate-limiting step in the growth of the database. As a result, these

databases have a limited scale and the genes are not selected in a systematic fashion.

An alternative approach is automated text mining: a method which involves automated information extraction by searching documents for text strings and analyzing their frequency and context. This approach has been used successfully in several instances for biological applications. In most cases, it has been applied to extract information about the relationships or interactions that proteins or genes have with one another, in the literature or by functional annotation.<sup>3-7</sup> Thus far, few publications have applied text-mining to examine the global relationships between genes and diseases. Perez-Iratxeta et al. automatically examined the GO (Gene Ontology) annotation of genes and their predicted chromosomal locations in order to identify genes linked to inherited disorders.<sup>8</sup>

To obtain a more global understanding of disease development, it would be valuable to incorporate information regarding all possible gene-disease relationships, including biochemical, physiological, pharmacological, epidemiological, as well as genetic. This information would enable comprehensive comparisons between large experimental datasets and existing knowledge in the literature. This would accomplish two things. First, it would serve to validate experiments by demonstrating that known responses occur as predicted. Second, it would rapidly highlight which genes are corroborated by the literature and which genes are novel in a given context. We have utilized a computational approach to literature mining to produce a

\* To whom correspondence should be addressed: jlabae@hms.harvard.edu.

comprehensive set of gene-disease relationships. In addition, we have developed a novel approach to assess the strength of each association based on the frequency of citation and co-citation. We applied this tool to help interpret the data from a large micro-array gene expression experiment comparing normal and cancerous breast tissue.

## Methods

**MedGene Database.** MedGene is a relational database, storing disease and gene information from NCBI, text mining results, statistical scores, and hyperlinks to the primary literature. MedGene has a web-based user interface for users to query the database (<http://hipseq.med.harvard.edu/MedGene/>).

**Text Mining Algorithms.** MeSH files were downloaded from the MeSH web site at NLM (National Library of Medicine) (<http://www.nlm.nih.gov/mesh/meshhome.html>) and human disease categories were selected. LocusLink files were downloaded from the LocusLink web site at NCBI (<http://www.ncbi.nlm.gov/LocusLink/>). Official/preferred gene symbol, official/preferred gene name, and gene alternative symbols and names, all relevant annotations and URLs for each LocusLink record, were collected. Gene search terms were used for literature searching and included all qualified gene names, gene symbols, and gene family terms. Primary gene keys, predominantly qualified gene family terms and gene official/preferred symbols, were used to index Medline records. If the official/preferred gene symbols did not meet the standards to be an index, then qualified gene official/preferred names were used. A local copy of Medline records (up to July, 2002) was pre-selected.

A JAVA module examined the MeSH terms and then indexed each Medline record with the appropriate disease terms. A separate JAVA module was used to examine the titles and abstracts for gene search terms and then to index the gene-related Medline records with the relevant primary gene key(s).

**Statistical Methods.** For every gene and disease pair, we counted records that were indexed for both gene and disease (double positive hits), for disease only (disease single hits), for gene only (gene single hits), and for neither gene nor disease (double negative hits) to generate a  $2 \times 2$  contingency table. On the basis of the contingency table framework, we applied different statistical methods to estimate the strength of gene-disease relationships and evaluated the results. These methods included chi-square analysis, Fisher's exact probabilities, relative risk of gene, and relative risk of disease<sup>10</sup> (<http://hipseq.med.harvard.edu/MedGene/>). In addition, we computed the "product of frequency", which is the product of the proportion of disease/gene double hits to disease single hits and the proportion of disease/gene double hits to gene single hits. To obtain a normal distribution, we transformed all the statistical scores using the natural logarithm. We selected the log of the product of frequency (LPF) to validate MedGene and to use for the analysis with the micro-array data. Spearman rank-correlation coefficients were used to assess the linear relationship between LPF and micro-array fold change in expression level.

**Global Analysis.** Diseases with at least 50 related genes were selected for clustering analysis, and the LPF scores were normalized with total score for each disease. Hierarchical clustering was done with the "Cluster" software and the clustering result was visualized using "TreeView" (<http://rana.lbl.gov/EisenSoftware.htm>).

**Breast Tissue Micro-Arrays.** Eighty-nine breast cancer samples (79% ER-positive) and 7 normal breast tissue samples were selected from the Harvard Breast SPORC frozen tissue repository and were representative of the spectrum of histological types, grades, and hormone receptor immunophenotypes of breast cancer. Biotinylated cRNA, generated from the total RNA extracted from the bulk tumor, was hybridized to Affymetrix U95A oligo-nucleotide micro-arrays. These micro-arrays consist of 12 400 probes, which represent approximately 9000 genes. Raw expression values were obtained using GENE-CHIP software from Affymetrix, and then further analyzed using the DNA-Chip Analyzer (dChip) custom software.

## Results

**Automated Indexing of Medline Records by Disease and Gene.** To study the gene-disease associations in the literature, we first compiled complete lists for human diseases and human genes. To index all Medline records that were relevant to human diseases, the Medical Subject Heading (MeSH) Index of Medline records was utilized. MeSH is a controlled medical vocabulary from the National Library of Medicine and consists of a set of terms or subject headings that are arranged in both an alphabetic and an hierarchical structure. Medline records are reviewed manually and MeSH terms are added to each with software assistance.<sup>9,10</sup> Twenty-three human disease category headings along with all of their child terms (see the Supporting Information, Supplemental Table 1, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table1.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table1.html)) were selected from the 2002 MeSH Index creating a list of 4033 human diseases.

No index comparable to the MeSH index exists for genes, and thus, it was necessary to apply a string search algorithm for gene names or symbols found in Medline text. A complete list of genes, gene names, gene symbols, and frequently used synonyms were collected from the LocusLink database at NCBI,<sup>11,12</sup> which contains 53 259 independent records keyed by an official gene symbol or name (June 18<sup>th</sup>, 2002). For the purposes of this study, no distinction was made between genes and their gene products. Authors often use the same name for both, differentiating the two only by the use of italics, if at all. For the intended use of this study, this lack of distinction is unlikely to have a large effect and may in fact be beneficial.

Initial attempts to search the literature using these lists revealed several sources of false positives and false negatives (Table 1). False positives primarily arose when the searched term had other meanings, whereas false negatives arose from syntax discrepancies necessitating the development of filters to reduce these errors. The syntax issues were readily handled by including alternate syntax forms in the search terms. The false positive cases, caused by duplicative and unrelated meanings for the terms, were more difficult to manage. Where possible, case sensitive string mapping reduced inappropriate citations. In many cases, however, this was not sufficient and the terms had to be eliminated entirely, thereby reducing the false positive rate but unavoidably under-representing some genes.

For the purposes of data tracking, a primary gene key was selected to represent all synonyms that correspond to each gene. Medline records were indexed with a primary gene key when any synonym for that key was found in the title or abstract. Case-insensitive string mapping was used for all searches except as noted above. No additional weight was



Table 1. Systematic Sources of False Positives and False Negatives in Unfiltered Data\*

source of error	error type	example	filter solution
gene symbol/name is not unique	false positive	MAG—myelin associated glycoprotein MAG—malignancy-associated protein	eliminate this term
gene symbol is unrelated abbreviation	false positive	PA—pallid homologue (mouse), pallidin (also abbrev. for Pennsylvania)	eliminate this term
gene symbol/name has language meaning	false positive	WAS—Wiskott-Aldrich Syndrome (also the word "was")	case-sensitive string search
nonstandard syntax	false negative	BAG-1 instead of BAG1	add dash term
unofficial gene name/symbol	false negative	P53 instead of TP53	add all gene nicknames
nonspecified gene name	false negative	estrogen receptor instead of Estrogen receptor 1	add family stem term

\* In preliminary studies, Medline was searched for co-occurrence of genes and diseases and the resulting output was evaluated to identify error sources that were amenable to global filters. Each error source is categorized by the type of error it causes: false positives are suggested relationships that are not real and false negatives are real relationships that are underrepresented. The filter solutions used are indicated. Note that in some cases, the filter solution itself introduces error. In general, error rates maximized sensitivity, even at the expense of specificity if needed.

added for multiple occurrences of a term or the co-occurrence of multiple synonyms for the same gene key.

Medline records were searched with all qualified gene identifiers, such as the official/preferred gene symbol, the official/preferred gene name, all gene nicknames and all syntax variants. In situations where there are several members of a gene family or splice variants, some authors prefer to use a shortened gene family name, e.g., estrogen receptor, instead of estrogen receptor 1 (*ESR1*), creating a source of false negatives. For this reason, gene family stem terms were created for all genes that have an alpha or numerical suffix (e.g., *IL2RA*, *TGFB*, *ESR1*, etc.) and then used to search the literature. The family stem terms were handled separately from the specific gene names so that it would be clear when linkages were made to the gene family versus a specific member in that family.

To improve performance and accuracy, some pre-selection was applied to the records that were scanned. First, review articles were eliminated to avoid redundant treatment of citations. Second, non-English journals were removed because the natural language filters were only relevant to English publications. Finally, journals unlikely to contain primary data about gene-disease relationships were also removed (e.g., *Int. J. Health Educ.*, *Bedside Nurse*, and *J. Health Econ.*). Together, these filters reduced the 12 198 221 Medline publications (July 2002) by 37%.

**Ranking the Relative Strengths of Gene-Disease Associations.** In total, there were 618 708 gene-disease co-citations, in which 16% (8297) of all studied genes had been associated to a disease and 96% (3875) of all diseases had been associated to at least one gene. To rank the relative strengths of gene-disease relationships, we tested several different statistical methods and examined the results. With the exception of the relative risk estimates, the methods provided similar results with respect to the rank order of the gene-disease association strengths. However, after comparing the results to other databases and after consulting disease experts, the log of the product of frequency (LPF) was selected for further analysis because it gave the best results overall.

**Validation of MedGene.** In developing this tool, it was important to minimize the number of missed genes (false negatives) and misclassified genes (false positives). However, in situations when these goals were in conflict, inclusiveness was prioritized. To determine the false negative rate in MedGene, breast cancer was used as a test case because it was associated with more genes than any other human disease and because

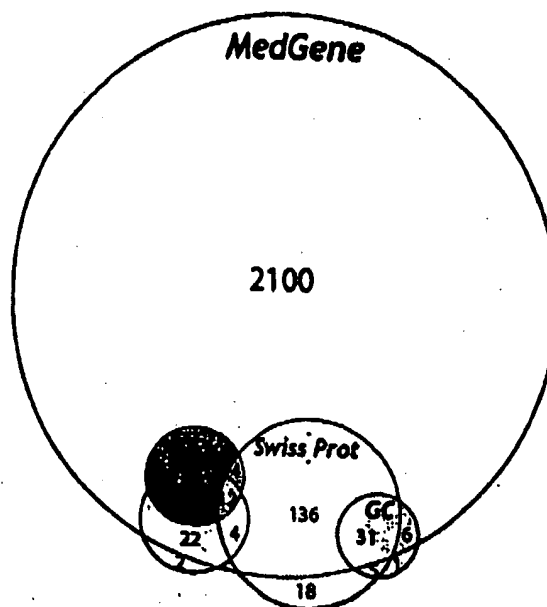


Figure 1. Estimation of the false negative rate by comparison with hand-curated databases. The breast cancer-related genes identified by MedGene were compared with those listed in several other databases including the Tumor Gene Database (TGD),<sup>2</sup> the Breast Cancer Gene Database (BCG),<sup>1</sup> GeneCards (GC)<sup>17</sup> and Swissprot.<sup>18</sup> Genes were considered false negatives if they were represented in at least one of these other databases and not in MedGene and their link to breast cancer was supported by at least one literature reference. All literature references were verified by manual review to confirm their validity. The number of genes in each database or shared by more than one database is indicated. The false negative rate was calculated by genes missed at MedGene (26)/total number of nonoverlapping genes in other databases (285).

there were several public databases that link genes to breast cancer. We compared the list of breast cancer-related genes from MedGene to these databases, illustrated in Figure 1. Among the 285 distinct breast cancer-related genes that were supported by at least one literature citation in these hand-curated databases, 26 were absent from MedGene, suggesting a false negative rate of approximately 9%. To determine why these were missed, all literature references for these genes (80



papers) were reviewed manually (see the Supporting Information, Supplemental Table 2, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table\\_2.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table_2.html)). Among these papers, most false negatives were caused by nonstandard gene terms or gene terms eliminated by our specificity filters. Few genes were missed because they were only mentioned in review papers (0.4%) or they appeared only in the body of the manuscript but not the abstract or title (1.1%). Of note, MedGene identified approximately 2000 additional breast cancer-related genes not listed in any other database.

To assess the false positive error rate, two complementary approaches were used: a detailed analysis of one disease and a global examination of 1000 diseases. The detailed approach examined the false positive error rate and its sources, whereas the global approach tested whether the overall results made biomedical sense.

Using the LPF, 1467 genes related to prostate cancer were assembled in rank order. We then retrieved approximately 300 Medline records each for the highest ranked 100 and the lowest ranked 200 genes and manually reviewed the titles and abstracts to determine the verity of the association. Nearly 80% of the highest ranked 100 genes fell into one of the five categories that reflect meaningful gene-disease relationships (see the Supporting Information, Supplemental Table 3, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table\\_3.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table_3.html)). Among the lowest ranked 200 genes, approximately 70% reflected true relationships. Of the 600 records reviewed, there were only two in which the association between the gene and the disease was described as negative. Both were genes with very low scores. In both cases, the authors did not argue the absence of any relationship, but rather that a particular feature of the gene or protein was not shown to be related to human prostate cancer.<sup>13,14</sup>

The coincidence of some gene symbols with medical abbreviations, chemical abbreviations and biological abbreviations resulted in most of the false positives (see the Supporting Information, Supplemental Table 4, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table\\_4.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table_4.html)), emphasizing the importance of the filters that were added in the search algorithm (Table 1). Without the filters, the false positive rate more than doubled, and the false negative rate rose dramatically (data not shown). For example, among the papers about breast cancer, there were only 12 Medline records that referred to *ESR1* and 10 to *ESR2*, whereas almost 2000 papers mentioned estrogen receptor without specifying *ESR1* or *ESR2*; this latter group was detected by the family term filter.

To further validate these results, a global analysis of the gene-disease relationships described by MedGene was performed. For this experiment, it was reasoned that the more closely related the diseases are to one another, the more they will be related to the same gene sets. Thus, if the relationships defined by MedGene accurately reflected the literature, then an unsupervised hierarchical clustering of the gene data should group diseases in a manner consistent with common medical thinking. Conversely, if the clustered diseases do not make sense biologically or medically, it may reflect excessive false positives, false negatives, or inappropriate scoring of the data.

To execute this experiment, the gene sets and the corresponding LPF values for 1000 randomly selected diseases (each with at least 50 gene relationships) were used as a dataset for clustering the diseases. A review of the results showed that the resulting disease clusters were indeed logical based upon common medical knowledge (see the Supporting Information,

Supplemental Figure 1, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Figure\\_1.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Figure_1.html)). For example, in one such cluster shown in Figure 2, diabetes and its complications grouped together and were also closely linked to diseases associated with starvation states.

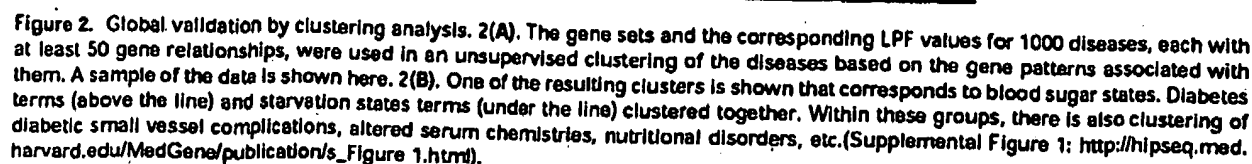
The number of genes associated with a given disease can be estimated by adjusting the MedGene number up by the false negative rate (~9%) and down by the false positive rate (~26% on average). Using this, the average disease has  $103.7 \pm 45.3$  (mean  $\pm$  s.d.) genes associated with it, although the range is quite broad with 2359 genes related to breast cancer, 2122 genes related to lung cancer and no genes related to a number of diseases.

**Applying MedGene to the Analysis of Large Datasets.** Access to a comprehensive summary of the genes linked to human diseases provided an opportunity to analyze data obtained from a high-throughput experiment. We compared the MedGene breast cancer gene list to a gene expression data set generated from a micro-array analysis comparing breast cancer and normal breast tissue samples. Micro-array analysis identified 2286 genes that had greater than a 1-fold difference in mean expression level between breast cancer samples and normal breast samples. Using MedGene, we sorted the 2286 genes into four classes: 555 genes directly linked to breast cancer in the literature by gene term search (first-degree association by gene name); 328 genes directly linked by family term search (first-degree association by family term); 1021 genes linked to breast cancer only through other breast cancer genes (second-degree association); and 505 genes not previously associated with breast cancer. (See the Supporting Information, Supplemental Figure 2, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Figure\\_2.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Figure_2.html).) Among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease. Among the remaining 38 genes, 9 had been related to other cancers, specifically esophageal, colon, uterine, skin, and cervix.

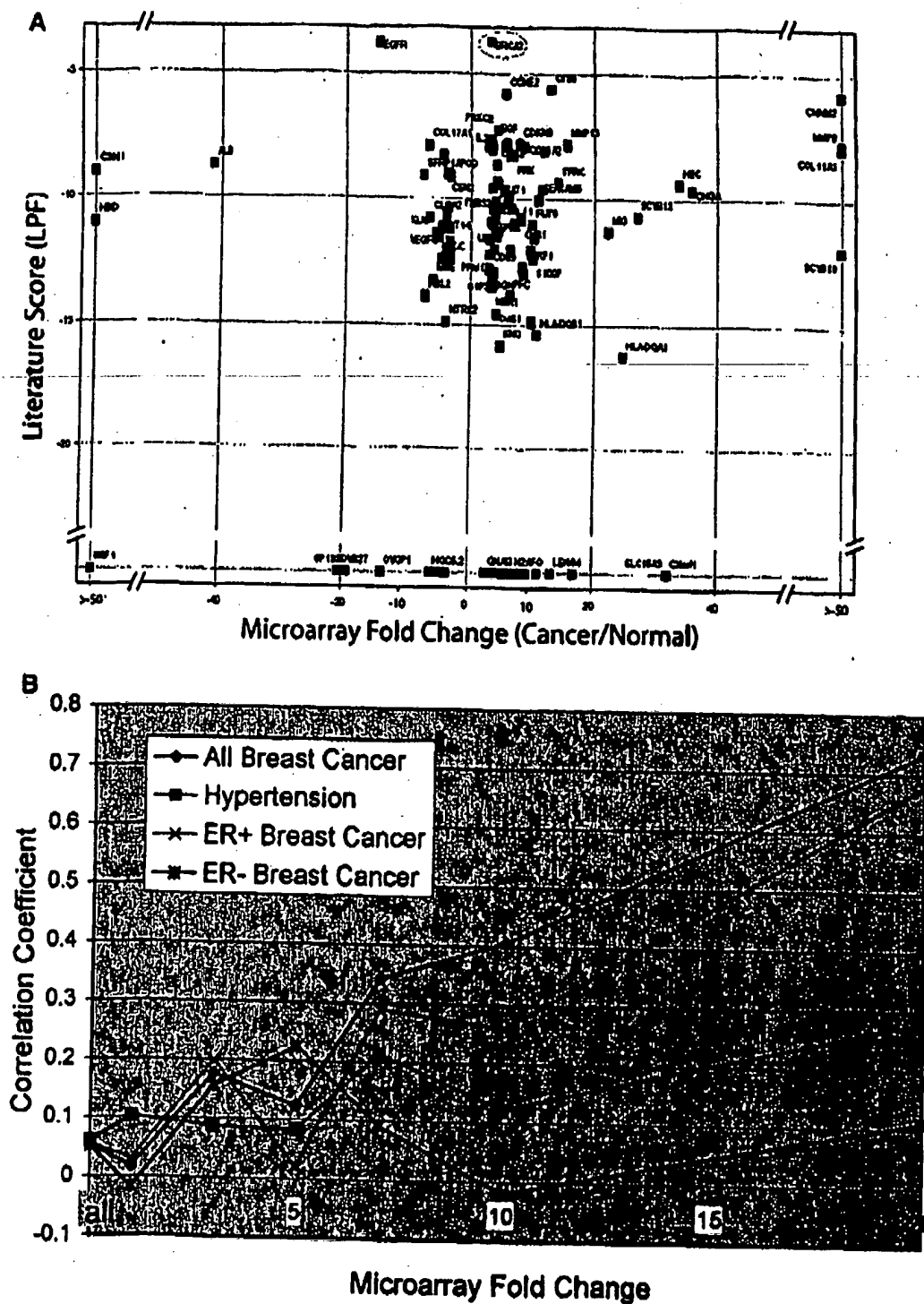
To determine whether the genes highlighted by the micro-array analysis were more likely to have been previously linked to breast cancer in the literature, we created a two-dimensional plot of the fold change of expression level between breast cancer and normal tissue versus the literature score (LPF) (Figure 3A). There was a broad spread of expression changes among the genes directly linked to breast cancer ranging from less than 1-fold change (68%) to over 40-fold (0.3%). Notably, the majority of genes with greater than 10-fold expression changes were linked to breast cancer by first-degree association.

Among all 754 genes directly linked to breast cancer in the literature, there was no correlation between LPF and micro-array fold change ( $r = 0.018$ ,  $p$ -value = 0.62). However, when we stratified the analysis based on the magnitude of the fold change, we observed an increasing trend in correlation (Figure 3B) suggesting that genes with a more substantial change in expression level were more likely to have a stronger association in the literature. For genes that had 10-fold change or more in expression level, the correlation increased to 0.41 ( $p$ -value = 0.05).

When we evaluated the micro-array data separately for ER positive and ER negative tumors, the trend in correlation between fold change and literature score was highly dependent on estrogen receptor status. Interestingly, there was a similar trend in correlation for ER positive tumors, but no trend in correlation for ER negative tumors.



disease unrelated to breast cancer. As expected, we did not observe an increasing trend in correlation for hypertension.



**Figure 3.** Relationship between literature score and functional data for breast cancer. **3A.** The data from an expression analysis of samples for breast tumors and normal breast tissue were analyzed to indicate the fold difference of expression level between breast tumor and normal sample (cutoff  $\geq 3$ -fold change). The fold changes were plotted against the literature score for the same gene set. Green dots represent first-degree association by gene search, blue dots represent first-degree association by family search and red dots represent no-association. Some well-studied genes, such as BRCA2 (pink circle), are not reflected by a substantial difference in expression level. Furthermore, the majority of genes that have no association with breast cancer in the literature had less than 10-fold expression changes (shaded area). **3B.** The Spearman rank-correlation coefficients between literature score (LPF) and the fold change of expression level between tumor and normal breast samples (y-axis) in relation to the amount of fold change of expression level (x-axis). Gene rank lists were generated for breast cancer (blue) and hypertension (pink). Correlations were also computed between the breast cancer gene LPF scores and fold change expression data among estrogen receptor positive tumors only (light blue) and estrogen receptor negative tumors only (purple).

Table 2. Top 25 Genes Related to Selected Human Diseases\*

breast neoplasms	hypertension	rheumatoid arthritis	bipolar disorder	atherosclerosis
estrogen receptor	<i>REN</i>	<i>RA</i>	<i>ERDA1</i>	apolipoprotein
<i>PCR</i>	<i>DBP</i>	<i>TNFRSF10A</i>	<i>SNAP29</i>	<i>APOE</i>
<i>ERBB2</i>	<i>LEP</i>	<i>CRP</i>	<i>PFKL</i>	<i>LDLR</i>
<i>BRCA1</i>	<i>AGT</i>	<i>AS</i>	<i>DRD2</i>	<i>ELN</i>
<i>BRCA2</i>	<i>INS</i>	<i>ESR1</i>	<i>TRH</i>	<i>ARG1</i>
<i>EGFR</i>	kallikrein	<i>HLA-DRB1</i>	<i>IMPA2</i>	<i>APOB</i>
<i>CYP19</i>	<i>ACE</i>	<i>DR1</i>	<i>HTR3A</i>	<i>APOA1</i>
<i>TFF1</i>	endothelin	interleukin	<i>DRD3</i>	<i>MSR1</i>
<i>PSEN2</i>	<i>S100A6</i>	<i>TNF</i>	<i>REM</i>	<i>LPL</i>
<i>TP53</i>	<i>BDK</i>	<i>IL6</i>	<i>KCNN3</i>	<i>PON1</i>
<i>CES3</i>	<i>DIAPH</i>	collagen	<i>DRD4</i>	plasminogen
<i>CEACAM5</i>	<i>SAR1</i>	<i>IL1A</i>	<i>HTR2C</i>	activator inhibitor
<i>ERBB3</i>	<i>PIH</i>	<i>ACR</i>	<i>RELN</i>	<i>PLG</i>
cydin	<i>CD59</i>	<i>TNFRSF12</i>	<i>DBH</i>	vascular cell
<i>COX5A</i>	<i>ALB</i>	<i>IL2</i>	<i>MAOA</i>	adhesion molecule
cathepsin	<i>CYP11B2</i>	<i>CHI3L1</i>	<i>COMT</i>	<i>ATOH1</i>
<i>ERBB4</i>	<i>MAT2B</i>	<i>IL8</i>	<i>HTR2A</i>	<i>VWF</i>
<i>TRAM</i>	angiotensin receptor	interleukin 1 matrix metalloproteinase	<i>SYNJ1</i>	<i>INS</i>
<i>CCND1</i>	<i>AGTR2</i>	interferon	<i>INPP1</i>	<i>ARG2</i>
<i>EGF</i>	<i>NPPA</i>	<i>CD68</i>	<i>NEDD4L</i>	<i>ABCA1</i>
<i>MUC1</i>	<i>LVM</i>	<i>IL4</i>	<i>FRA13C</i>	<i>OLR1</i>
insulin-like	<i>DBH</i>	<i>IL17</i>	transducer of	collagen
<i>BCL2</i>	<i>NPY</i>	<i>MMP3</i>	<i>ERBB2</i>	<i>MCP</i>
mucin	<i>POMC</i>	<i>SIL</i>	<i>BAIAP3</i>	lipoprotein
<i>FGF3</i>	neuropeptide		<i>ATP1B3</i>	<i>APOA2</i>
			<i>DRD5</i>	intercellular adhesion molecule
				<i>RAB27A</i>

\* MedGene results for the top 25 genes associated with breast neoplasms, hypertension, rheumatoid arthritis, bipolar disorder, and atherosclerosis, respectively, ranked by LPP scores. The hyperlink to all the papers co-citing the gene and the disease is available at MedGene website (<http://hipseq.med.harvard.edu/MedGene/>).

## Discussion

The Human Genome Project heralded a new era in biological research where the emphasis on understanding specific pathways has expanded to global studies of genomic organization and biological systems. High-throughput technologies can provide novel insight into comprehensive biological function but also introduces new challenges. The utility of these technologies is limited to the ability to generate, analyze, and interpret large gene lists. MedGene, a relational database derived by mining the information in Medline, was created to address this need. MedGene users can query for a rank-ordered list of human gene-disease relationships (Table 2) for one or more diseases. Each entry is hyperlinked to the original papers supporting each association and to other relevant databases.

MedGene is an innovative extension of previous text mining approaches. Perez-Iratxeta et al. used the GO annotation and their chromosomal locations to predict genes that may contribute to inherited disorders.<sup>8</sup> MedGene takes a broader view and includes all diseases and all possible gene-disease relationships. Furthermore, MedGene utilizes co-citation to indicate a relationship rather than GO annotation, which is limited to the subset of genes that have GO annotation. Our approach is complementary to that taken by Chaussabel and Sher, who used the frequency of co-cited terms to cluster genes into a hierarchy of gene-gene relationships.<sup>9</sup>

A unique aspect of this tool is the ability to assess the relative strengths of gene-disease relationships based on the frequency of both co-citation and single citation. This presupposes that most co-citations describe a positive association, often referred to as publication bias<sup>15</sup> and is supported by our observations

that negative associations are rare (Supplemental Table 3: [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table3.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table3.html)). Of course, relationships established by frequency of co-citation do not necessarily represent a true biological link; however, it is strong evidence to support a true relationship.

Another important feature of MedGene is the implementation of software filters that substantially reduced the error rate. We estimate that less than 10% of all associations were missed and at least 70% of even the weakest associations were real. For this study, all of the filters that we applied were general ones, e.g., expanding the list of all gene names to address the different syntax forms used by different journals, eliminating gene names that correspond to common English words, etc. The majority of the remaining search term ambiguities were idiosyncratic and difficult to identify systematically without causing a significant rise in false negatives. Alternative approaches, such as the examination of the nearest neighbor terms, need to be considered to further reduce the false positive rate.

It is not uncommon to see expression changes in microarray experiments as small as 2-fold reported in the literature. Even when these expression changes are statistically significant, it is not always clear if they are biologically meaningful. When comparing expression levels of disease to normal tissue, one expects an enrichment of known disease-related genes to appear in the altered expression group. MedGene provided a unique opportunity to test this notion in the context of existing knowledge on a novel breast cancer microarray dataset. For genes displaying a 5-fold change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. This

**Table 3.** Genes with Large Expression Changes in ER- but Not in ER+ Breast Tumors

gene symbol	fold change (ER+)	fold change (ER-)
KRTHB1	1.0	610.8
BRS3	1.2	89.4
DKK1	1.2	69.8
ZIC1	1.9	59.6
TLR1	1.0	38.5
KIAA0680	2.6	33.2
CDKN3	1.0	30.6
EBI2	4.0	27.9
GZMB	3.8	21.9
STK18	4.7	18.6
GPR49	1.0	14.6
MYO10	1.6	14.4
LAD1	-1.0	13.5
POLE2	4.2	13.0
HMG4	4.4	12.9
BCL2L11	-1.2	12.3
LRP8	2.9	12.2
CCNB2	1.0	11.8
CCNE2	4.0	11.6
FGF	-4.3	11.1
KNL6	2.9	10.9
HIF5	3.0	10.2
SERPINH2	4.6	10.2
YAP1	1.0	10.0
LPHB	-1.3	-10.4
TCEA2	-1.1	-10.8
TFF1	1.3	-11.4
COL17A1	-4.1	-15.7
POP5	1.1	-16.2
BPAG1	-4.6	-22.3
PDZK1	-1.1	-36.8
VECF	-2.8	-51.5
MUC6	-1.4	-64.9
SERPINA5	-1.0	-83.1
MEIS1	-1.6	-85.9
CA12	2.4	-150.3

Table 3. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative, but not ER positive breast tumors. All of these genes have either never been co-cited with breast cancer or have a weak association except those marked with an \*.

reflects the many genes whose role in breast cancer may not involve large changes in expression in sporadic tumors (e.g., *BRCA1* and *BRCA2*) and genes whose modest changes in expression may be unrelated to the disease. Strikingly, among genes with a 10-fold change or more in expression level, there was a strong and significant correlation between expression level and a published role in the disease, providing the first global validation of the micro-array approach to identifying disease-specific genes.

The results derived from MedGene have two implications. First, a careful hunt for corroborating evidence of a role in breast cancer should precede any further study of genes with less than 5-fold expression level changes. Second, any genes with 10-fold changes or more are likely to be related to breast cancer and warrant attention. It is likely that this threshold will change depending on the disease as well as the experiment.

Interestingly, the observed correlation was only found among ER-positive tumors, not ER-negative. This may reflect a bias in the literature to study the more prevalent type of tumor in the population. Furthermore, this emphasizes that caution must be taken when interpreting experiments that may contain subpopulations that behave very differently. The MedGene approach identified a set of relatively understudied, yet highly expressed genes in ER-negative tumors that are worthy of further examination (Table 3).

In conclusion, we have developed an automated method of summarizing and organizing the vast biomedical literature. To our knowledge, the resulting database is the most comprehensive and accurate of its kind. By generating a score that reflects the strength of the association, it provides an important tool for the rapid and flexible analysis of large datasets from various high-throughput screening experiments. Furthermore, it can be used for selecting subsets of genes for functional studies, for building disease-specific arrays, for looking at genes common to multiple diseases and various other high-throughput applications. In the future, it will be possible to enhance the utility of the MedGene database by building links between genes and other MeSH terms as well as other biological processes and concepts, such as cell division and responses to small molecules.

**Acknowledgment.** We would like to thank P. Braun, L. Garraway, J. Pearlberg, and other members of our institute for helpful discussion. Many thanks to the NLM (National Library of Medicine) for licensing of MEDLINE and the annotation effort of adding MeSH Indexes for MEDLINE abstracts. This work was funded by grants from the Breast Cancer Research Foundation and an NHLBI PGA Grant (Vol HL66582-02).

**Supporting Information Available:** Twenty-three human disease category headings along with all of their child terms selected from the 2002 MeSH Index (Supplemental Table 1); analysis of the causes of false negatives in MedGene (Supplemental Table 2); meaningful gene-disease relationships found in MedGene (Supplemental Table 3); causes for incorrect assignment of gene indexes (Supplemental Table 4); a review of the results, showing that the resulting disease clusters were indeed logical (Supplemental Figure 1); and a review of the results showing that among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease (Supplemental Figure 2). This material is available free of charge via the Internet at <http://pubs.acs.org> and at the web sites mentioned in the text.

## References

- (1) Baasiri, R. A.; Glasser, S. R.; Steffen, D. L.; Wheeler, D. A. *Oncogene* 1999, 18, 7958-7965.
- (2) Steffen, D. L.; Levine, A. E.; Yarus, S.; Baasiri, R. A.; Wheeler, D. A. *Bioinformatics* 2000, 16, 639-649.
- (3) Marcotte, E. M.; Xenarios, I.; Eisenberg, D. *Bioinformatics* 2001, 17, 359-363.
- (4) Ono, T.; Hishigaki, H.; Tanigami, A.; Takagi, T. *Bioinformatics* 2001, 17, 155-161.
- (5) Jensen, T. K.; Laegreid, A.; Komorowski, J.; Hovig, E. *Nat. Genet.* 2001, 28, 21-28.
- (6) Chausabel, D.; Sher, A. *Genome Biol.* 2002, 3, RESEARCH0055.
- (7) Gibbons, F. D.; Roth, F. P. *Genome Res.* 2002, 12, 1574-1581.
- (8) Perez-Iratxeta, C.; Bork, P.; Andrade, M. A. *Nat. Genet.* 2002, 31, 316-319.
- (9) Funk, M. E.; Reid, C. A. *Bull. Med. Libr. Assoc.* 1983, 71, 176-183.
- (10) Humphrey, S. M.; Miller, N. E. *J. Am. Soc. Inf. Sci.* 1987, 38, 184-196.
- (11) Maglott, D. R.; Katz, K. S.; Scotte, H.; Pruitt, K. D. *Nucleic Acids Res.* 2000, 28, 126-128.
- (12) Pruitt, K. D.; Maglott, D. R. *Nucleic Acids Res.* 2001, 29, 137-140.
- (13) Wadelius, M.; Andersson, A. O.; Johansson, J. E.; Wadelius, C.; Rane, E. *Pharmacogenetics* 1999, 9, 333-340.
- (14) Adam, R. M.; Borer, J. G.; Williams, J.; Eastham, J. A.; Loughlin, K. R.; Freeman, M. R. *Endocrinology* 1998, 140, 5860-5875.
- (15) Montori, V. M.; Smieja, M.; Guyatt, G. H. *Mayo Clin. Proc.* 2000, 75, 1284-1288.
- (16) Denenberg, V. H. *Statistics Experimental Design for Behavioral and Biological Research*; Wiley-Liss: New York, 1978.
- (17) Rebhan, M.; Chalifa-Caspi, V.; Prilusky, J.; Lancet, D. *Trends Genet.* 1997, 13, 163.
- (18) Balroch, A.; Apweiler, R. *Nucleic Acids Res.* 2000, 28, 45-48, PR0340227

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**